Detecting anticipatory information in speech with signal chopping

Sam Tilsen
tilsen@cornell.edu
203 Morrill Hall
Ithaca, NY 14853

Abstract

Most analyses of articulatory processes in speech assume that word form-related changes in the state of the vocal tract have well-defined beginnings and ends. But how do we determine the precise moments in time when these beginnings and ends occur? More specifically, when should we expect information related to the sound categories of a word to be present in acoustic and articulatory signals? The framework of Articulatory Phonology / Task Dynamics predicts that the earliest time such information becomes available is when the first articulatory gesture of a word becomes active, which closely corresponds to when a movement is initiated. Alternatively, a recent extension of the Articulatory Phonology model holds that gestures may have an influence on the state of the vocal tract after they have been retrieved from memory, but before they become active and before canonical movement initiation. This paper presents evidence that indeed, anticipatory information is available much earlier than is typically assumed: the identity of a syllable onset gesture can be predicted from articulatory and acoustic data quite early, in some cases nearly half a second before movement initiation. Likewise, the identity of a coda gesture can be predicted during the period of time typically associated with an onset consonant. These findings were obtained with a novel analysis method called *signal chopping* which was paired with deep neural network based classification. In this approach articulatory and acoustic signals are systematically truncated in space and time, and a network training/test procedure is repeated on the chopped signals. By analyzing the effects of chopping on classification accuracy, gesture-specific information can be spatiotemporally localized.

Keywords: speech planning; production; articulation; Articulatory Phonology; neural networks, deep learning; information; signal chopping

1. Introduction

Consider separate utterances of CVC word forms, such as *pop* and *pot*. When in time do we expect acoustic or articulatory signals associated with these words to differ? A standard analysis of such forms in the Articulatory Phonology (AP) / Task Dynamics (TD) framework holds that the bilabial constriction gesture associated with coda /p/ and the alveolar constriction gesture associated with coda /t/ first become active some time after the onset consonantal gesture deactivates. Thus, the coda becomes active well after the release of the onset constriction. There should be no information regarding the identity of the coda gesture before that time. Similarly, there should be no information regarding the identity of the onset in words like *pop* and *top*, until the onset consonantal gestures have become active. Analyses of experimental data presented here show that, contrary to these predictions, anticipatory information regarding coda gestures is often present during onsets, and anticipatory information regarding onset gestures can be available well before canonical movement initiation. The anticipatory patterns are predicted by a revised model of how gestures influence the state of the vocal tract. A number of fundamental issues are addressed in this paper, including what it means for a gesture to be initiated, and how we can use information in articulatory and acoustic signals to identify gestural categories in speech.

The paper is organized as follows. In section (1.1) predictions of the standard AP/TD model regarding anticipatory information are discussed. In section (1.2) empirical evidence from previous studies which challenges the AP/TD model is presented. Section (1.3) describes a revision to AP/TD which allows for anticipatory information to be present after gestural retrieval and before canonical activation. Section (1.4) provides a conceptual description of the signal chopping method that is used in this paper. Section (1.5) delineates the experimental hypotheses and predictions. Section 2 describes the experiment and data analysis methods in detail, and sections 3 presents the results. Section 4 discusses how the results bear on our understanding of information in speech signals and proposes several directions for future research.

1.1 Predictions of standard Articulatory Phonology

Articulatory Phonology (AP) / Task Dynamics (TD) does not generate extensive anticipatory effects. To see why, lets consider the AP gestural score and the corresponding tract variable changes generated by the TD model (Browman & Goldstein, 1992; Saltzman & Munhall, 1989; Tilsen, 2019). As illustrated in Fig. 1, gestural activation intervals in the score are periods of time in which a force acts upon a tract variable system, driving it toward a target state associated with the gesture. For instance, lip aperture (LA) is a tract variable system whose state is a degree of opening between the lips; a bilabial closure gesture |LAB clo| drives the LA system to a state in which the lips are closed. A crucial point to make is that, in standard implementations of the AP/TD model, a gesture cannot have an influence on the vocal tract before it becomes active. This leads to two predictions. The *onset prediction* is that in a period of time preceding the initiation of the onset constriction gesture (i.e. the pre-onset epoch), there should be no information available in acoustic or articulatory signals which can be used to predict the identity of the onset. The *coda prediction* is that in a period of time preceding the initiation of the coda constriction gesture (i.e. the pre-coda epoch), there should be no information which can be used to predict the identity of the coda.

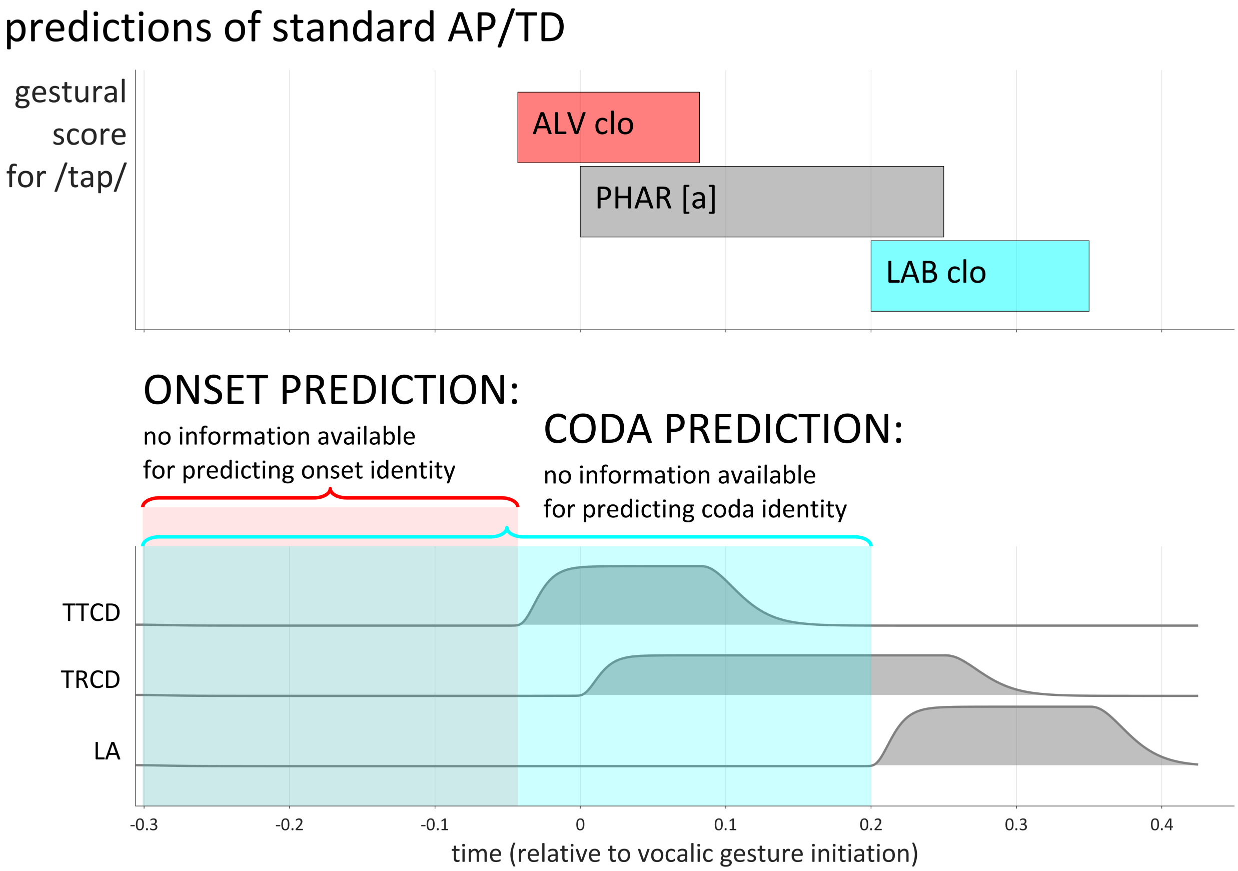


Fig. 1. Predictions of the standard AP/TD model regarding the availability of information to identify the onset and coda of a CVC syllable. (Top) Gestural score for the word *top*, with an alveolar closure gesture, a pharyngeal constriction for the vowel [a], and a bilabial closure gesture. (Bottom) Tract variable changes driven by gestural activation. TTCD: tongue tip constriction degree; TRCD: tongue root constriction degree; LA: lip aperture.

Both of the predictions can be understood to follow from two aspects of the AP/TD framework, which are (I) the assumption of parameter invariance and (II) hypothesized onset/coda timing asymmetries. First, regarding parameter invariance, in AP/TD each gesture is associated with a target value of a tract variable, as well as a stiffness which describes the strength of the influence of that gesture on the tract variable. The parameters of gestures are typically assumed not to vary as a function of the phonological environment (Browman & Goldstein, 1992; Saltzman & Munhall, 1989). For example, the target and stiffness of a |LAB clo| gesture associated with the onset /p/ in *pot* are not contingent on the identity of the coda and hence are identical to the parameters of |LAB clo| in *pod, pot, pock, etc*. Parameter invariance is desirable because it leads to a more parsimonious account of the information which is retained in lexical memory: with a fairly small set of gestural parameters, surface variation can be generated by blending the forces of temporally overlapping gestures and by allowing for biomechanical interactions between articulators.

However, phonological context-dependence of gestural parameters is not strictly forbidden by AP and may be necessary for empirical adequacy. Assimilatory sound changes between neighboring segments are common, and these suggest that when gestures overlap in time, speakers may learn context-specific gestural parameters. Exactly how many parameters this could involve depends on how the contexts are defined. For example, there are approximately 121 word-initial CV combinations in English (when counting only oral constriction gestures and vocalic targets), and so one might allow for 121 different onset constriction gesture targets, one for each environment. Expanding the definition of contexts to include complex onsets and codas, and taking into account laryngeal gestures, nasal gestures, and vowel-associated labial gestures (lip rounding/retraction), the number of contexts would be proliferated by an order of magnitude. One context which is particularly relevant to the current study is the onset-coda context. Specifically, we might consider whether it is necessary to allow for the parameters of gestures in onset position to depend on the parameters of gestures in coda position, and vice versa. Such dependencies may be undesirable because they lead to further proliferation of information that must be retained in long-term memory.

Second, regarding onset/coda timing asymmetries, AP imposes temporal constraints on when gestures can begin to influence the vocal tract. Specifically, AP hypothesizes (i) that an onset consonantal constriction gesture in a CVC syllable becomes active in close temporal proximity to when the vocalic gesture becomes active, and (ii) that a coda constriction gesture becomes active substantially later than the vocalic gesture becomes active, well after the onset consonantal gesture has become inactive. These hypotheses are illustrated in Fig. 1 and are motivated by a substantial amount of empirical data (Browman & Goldstein, 1990; Byrd, 1996; Marin & Pouplier, 2010; Tilsen, 2017).

Under the assumptions of gestural target invariance and onset/coda timing asymmetry, the AP/TD model does not generate any articulatory or acoustic information regarding the place or manner of an onset consonant before the oral constriction gesture associated with the onset becomes active (Fig. 1, onset prediction). Nor does it generate information regarding the place or manner of a coda before the coda gesture becomes active (Fig. 1, coda prediction). It is important to recognize that if such information is shown to be present, it does not entail that the entire AP/TD framework is invalid, nor does it necessitate a context dependent proliferation of gestures. Such conclusions would be classic interpretive errors of the types discussed recently in (Mücke et al., 2020). Instead, this paper argues that the presence of anticipatory information can be understood with a substantial revision of AP/TD which has been developed in (Tilsen, 2018, 2019).

There are three important points to make here regarding the standard AP/TD model, in order to avoid potential confusion. The first involves the concept of *gestural initiation* and its correspondence with empirical estimates of movement onset. In a mathematical and theoretical sense, gestures in the AP/TD model are understood as systems which transition from an inactive state to an active state, i.e. from zero activation to maximal activation. The leftmost edges of the activation intervals in the gestural score of Fig. 1 correspond to beginnings of these transitions. As long as the transition from zero to maximal activation is relatively abrupt, there will be a close correspondence between the time of gestural initiation and time estimated from an empirical data using a standard criterion, such as 20% of maximal velocity. The AP/TD model, when optimized to fit empirical data, can be used assess this correspondence (see *Appendix: Correspondence between gestural initiation and empirical estimates*). Using empirical data from the current study, the optimized models show that gestural initiation precedes empirical estimates of movement onset by 13 ms on average, with a standard deviation of 3.1 ms. Note that gestural initiation is a theoretical event, derived from model optimization, while the empirical estimates are derived from tract variables using a velocity-based criterion. The calculation of the discrepancy is only possible because AP/TD provides a mathematically explicit model of how gestural activation influences tract variables. Crucially, the discrepancy is small enough that we can reasonably treat our empirical measures of movement onsets as estimates of gestural initiations.

The second important point to make is that AP/TD predicts that information regarding active gestures will be redundantly encoded, present in dimensions of articulatory signals which are not directly influenced by those gestures. Indeed, all oral constriction gestures influence multiple articulators and tract variables, and hence we do not expect that information for classifying gestures will be confined to any single articulator or tract variable. This is despite the fact that each gesture specifies a target for just one tract variable—e.g. a |LAB clo| gesture specifies a target lip aperture (LA). To see why, consider that many tract variables, and in particular all oral apertures/constriction degrees, are controlled with multiple articulators. For example, lip aperture (LA) is controlled with the jaw, lower lip, and upper lip; tongue tip constriction degree (TTCD) is controlled with the jaw and tongue blade. Notice that both tract variables, LA and TTCD, involve the jaw, and consider that the tongue blade and lower lip are biomechanically coupled to the jaw. Because of this shared mechanical relation with the jaw, an active gesture with a LA target will influence TTCD, and vice versa. All oral constriction gestures are expected to have such influences; the reader should consult (Saltzman & Munhall, 1989) for further explication. Thus we should have no a priori expectation that the information relevant for classifying a gesture such as |LAB clo| is confined to the articulators directly associated with lip aperture.

The third important point to make is that although some early descriptions of AP promoted the idea that onset consonantal and vocalic consonantal gestures in a CV syllable begin simultaneously, this claim has been revised in subsequent empirical and theoretical work. Currently the accepted generalization is that the beginnings of consonantal constriction and release gestures in a CV syllable are displaced in opposite directions from the beginning of the vocalic gesture (Nam, 2007; Tilsen, 2017). Empirical studies indicate that the initiation of the consonantal constriction gesture typically precedes the vocalic gestures onset by about 25-75 ms (Browman & Goldstein, 1995; Nam, 2007; Tilsen, 2017). Indeed, the timing pattern shown in Fig. 1 is the average pattern for the word *top* produced by a representative speaker in the current study. The timing pattern is hypothesized to arise from anti-phase coupling of planning oscillators associated with the constriction and release gestures, which are both in-phase coupled to the vocalic gesture planning oscillator.

1.2 Reasons to suspect that AP/TD undergenerates

There are several reasons to suspect that the AP/TD undergenerates. These are: (i) the existence of nonlocal anticipatory phonological patterns and speech errors, (ii) lexico-statistical patterns, and (iii) empirical evidence from past studies.

First, there are many examples of nonlocal assimilatory phonological patterns where (in segmental/featural terms) some feature of a segment must have a certain value when the same value of the feature appears later in a root or stem (Gafos, 1999; Hansson, 2001; Walker, 2011). Such patterns are "nonlocal" in that the interacting segments are not adjacent. A common example is root-internal sibilant harmony, illustrated schematically in Table 1. Consonant harmonies of this sort are often confined to roots or derived stems, and the interacting segments are typically similar. For example, the segments /s/ and /ʃ/ both involve a lingual constriction gesture which generates frication; they differ in that /s/ is [+anterior] and /ʃ/ is [-anterior], which on a gestural analysis amounts to whether the target place of articulation is alveolar or post-alveolar. Note that the target place of articulation is a gestural parameter which, under the assumption of gestural invariance, does not vary as a function of context.

Table 1. Pattern of well-formedness in sibilant harmony

sapas	*ʃapas
*sapaʃ	ʃapaʃ

The relevant question here is: how, on a diachronic timescale, do such patterns emerge? One not-so-plausible way is that for a particular speaker, some unknown process causes an instantaneous change in the long-term memories of segments, such that [+anterior] changes to [-anterior] in words which contain a following [-anterior], and vice versa. These innovations then spread through a population. Although anticipatory feature substitutions do seem to occur in speech errors, it is far-fetched to posit that substitutions are long-term memory phenomena. In other words, I may produce an anticipatory error such as *acoustic* [ʃ]*ignals show…*, but that does not mean that my long term memory of the first sound in the word *signals* has changed to [-anterior]. A more plausible explanation for the emergence of such patterns would be based on small, gradient changes and phonologization via hypocorrection (Ohala, 1993). For example, imagine there is some mechanism whereby the articulatory target for an active gesture can be perturbed to a gradient extent by the presence of another gesture in a word, even if that other gesture is not "active" in the conventional sense (Tilsen, 2019). The gradient perturbations may have subtle effects on the long term memory of an articulatory target, and repeated small perturbations may add up to sound changes that are re-analyzed as categorical shifts.

A second thread of evidence which suggests that standard AP undergenerates comes from lexical patterns in which there are statistical biases or constraints on the co-occurrence of features in a word form. For example, in English there are very few monosyllabic words of the form /sCVC/, where the immediately prevocalic onset and coda are homorganic, or both are nasal: word forms like *speb*, *skik*, and *snam* are conspicuously absent from the English lexicon with only a handful of exceptions involving alveolars (e.g. *stet*, *stat*), cf. (Clements & Keyser, 1983; Davis, 1989; Fudge, 1987). For another example, in Arabic triconsonantal roots, in addition to constraints against identity of the first two consonants (Greenberg, 1950; McCarthy, 1986), Pierrehumbert (1993) finds more general statistical biases against co-occurrence of consonantal features. Such statistical biases could arise from lexicalization of interactions between gestures which are not simultaneously active. It is more difficult to imagine how they could arise if only simultaneously active gestures can interact. Note that unlike non-local harmonies, statistical biases against co-occurrence of features are dissimilatory. Dissimilatory phenomena have been modeled with inhibitory interactions between gestures in target planning (Tilsen, 2019), but we omit discussion of such patterns in presenting the model below.

Third, there are various experimental studies that demonstrate anticipatory phenomena not readily modeled in standard AP/TD. Several studies have found evidence that speakers exert response-specific adjustments to the posture of the vocal tract prior to producing an utterance (Kawamoto et al., 2008; Krause & Kawamoto, 2019a, 2019b; Rastle & Davis, 2002; Tilsen et al., 2016). For example, Tilsen et al. (2016) used real-time MRI to examine vocal tract posture during a delay period prior to the production of /pa/, /ma/, /ta/, and /na/ syllables. By comparing prepared and unprepared response conditions, they found that many speakers adjusted vocal tract posture prior to initiating the response, but only when the upcoming response was known ahead of time. Importantly, these effects were partially assimilatory. For example, when a speaker knew they would produce /pa/ they would not necessarily close the lips completely during the delay period, yet lip aperture during this period was smaller than in the unprepared response condition. Such patterns show that an onset consonantal gesture can have an effect on the vocal tract well before the gesture "begins," in the standard AP sense. A different form of evidence for anticipatory phenomena comes from a study of reaction times in reading monosyllables. Cohen-Goldberg (2012) found that greater similarity between consonants in the onset and coda was correlated with slower reaction times (the effect size was approximately 20 ms over the full range of similarities). This effect suggests that there is an interaction between onset gestures and coda gestures, which manifests before a verbal response can be detected acoustically.

More generally, a variety of studies have found evidence for sub-categorical interactions which are non-local, i.e. between onsets and codas or between nonadjacent, heterosyllabic segments. Non-local interactions should be distinguished from local coarticulation arising from gestural overlap (cf. Tilsen, 2019 for a exposition of three definitions of locality). However, it is not always possible to distinguish between these without direct knowledge of gestural activation intervals. For example, various studies have observed assimilatory effects between formants of vowels in VCV sequences (Beddor et al., 2002; Öhman, 1967; Recasens, 1987), but it is unclear whether such effects are due to overlap of vocalic gestures between adjacent syllables, or whether they are a consequence of interactions between targets of contemporaneously planned but non-overlapping gestures. Coarticulatory effects between overlapping gestures can be “planned”, i.e. are partly under speaker control (Whalen, 1990), and so language- and speaker-specificity of such effects does not necessarily indicate that they arise from non-local interactions. More compelling examples of non-local interactions between vowels have been observed in effects on vowels from non-adjacent syllables (Grosvald, 2009; Magen, 1997), and in paradigms where a planned but non-executed vowel has an influence on a produced vowel (Tilsen, 2007, 2009). Moreover, observations of non-local interaction are not confined to vowels. For instance, Cohn (1990) observed differences in nasal airflow for nasal stops in onset position which were conditioned on the nasality of the coda. Hawkins & Nguyen (2004) found that the durations and formants of onset /l/ were different in syllables with voiced vs. voiceless codas. Below we present a model in which non-local interactions result from a planning mechanism which is distinct from gestural overlap and which operates before and after a gesture has been initiated/terminated.

1.3 The intentional planning model

In this section I present a brief overview of the intentional planning model, which can generate anticipatory patterns of the sort observed in the current study. This model is part of the Selection-coordination framework (Tilsen, 2013, 2014, 2016), an extension of Articulatory Phonology / Task Dynamics that incorporates a mechanism for competitive selection of sets of coordinated gestures. The intentional planning model has been described in detail in recent work (Tilsen, 2018, 2019), and the aim here is simply to provide a basic understanding of how the model allows for gestures which have not been initiated to influence the state of the vocal tract. To facilitate the exposition, a time course of production of the word *top* is shown in Fig. 2 and we refer to panels (i)-(viii), timepoints (t_1)-(t_6), and labels (A)-(E). The key ingredients of the model are described below.

Intentional planning fields. In the intentional planning model, a gesture is reconceptualized as a system which exerts a force on an intentional planning field. Each tract variable (TV), e.g. TTCD, LA, etc. is associated with a one-dimensional scalar field. The fields are topographically organized such that one end corresponds to a maximal value of some tract variable (e.g. maximal lip aperture) and the other end corresponds to a minimal value (e.g. minimal lip aperture). Each field experiences Gaussian force distributions from gestures and from a neutral attractor system. Examples of these forces on the LA field at the timepoints (t_1)-(t_6) are shown in (viii). Crucially, the current target of a given TV system is the centroid of the integrated field activation (vertical black lines in (viii)). This contrasts with the standard TD model in which the target of a TV system is a weighted average of the targets of active gestures. A hypothetical time evolution of the fields for TTCD, TRCD (tongue root constriction degree), and LA is shown with heatmaps in (v), along with the current centroid (i.e. the TV target, green lines). The vertical axes of the heat maps correspond to TV values and lighter colors indicate a greater degree of activation.

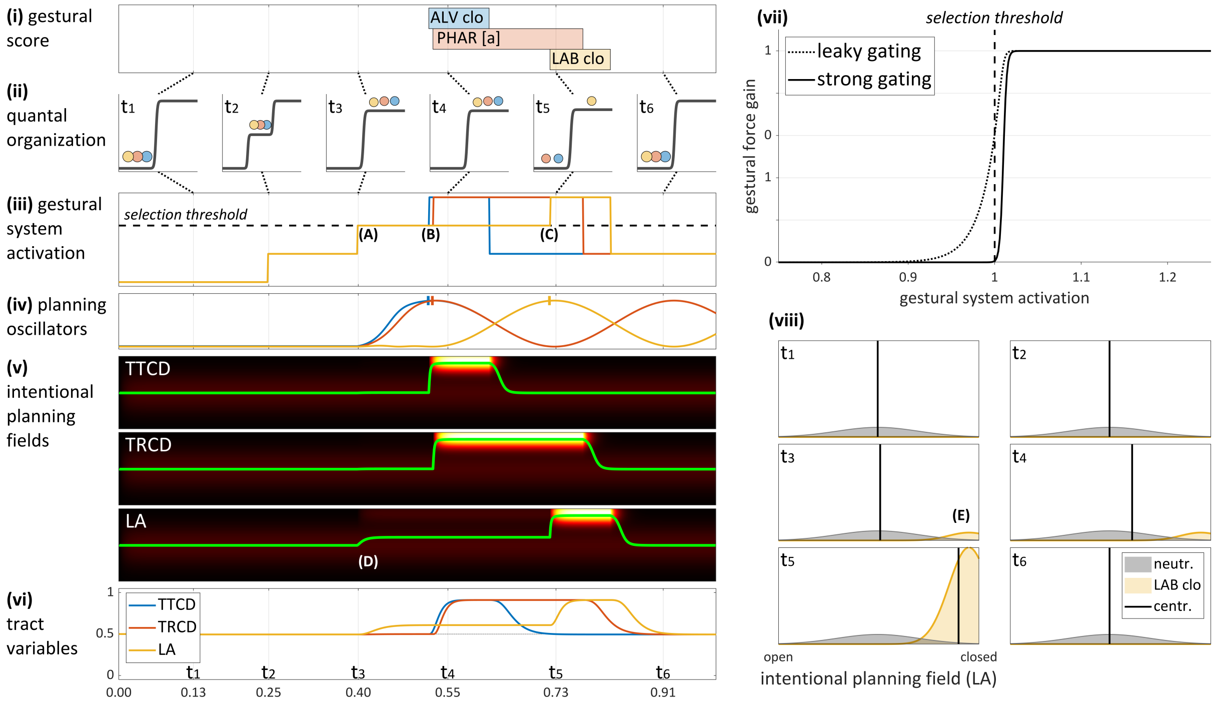


Fig. 2. Anticipatory effects in the intentional planning model. Prior to response initiation, gestural systems are organized in a quantal potential (ii). When the activations of the systems reach a selection threshold (iii), coupled planning oscillators (iv) determine when the gestures will be *initiated*, i.e. when they will exert strong forces on tract variables, which correspond to gestural activation intervals (i). Prior to this, subthreshold gestural activation may influence tract variables (vi) via the leaky gating mechanism (vii). The spatial and temporal effects of these influences are shown in heatmaps of activation in intentional planning fields (v). The result is that a gesture can influence the vocal tract state before it is initiated (D). See text for further detail.

Gestural systems and gestural planning systems. Each gesture in this model is conceptualized as a system which has a scalar state variable, referred to as *activation*, and shown schematically in (iii). In addition, each gesture is associated with a planning system that has oscillatory dynamics (iv). When a form is retrieved from the lexicon (or cued by a stimulus), the corresponding gestural systems transition to an excited state (t_2). In the post-retrieval excited state, the activation values of the gestural systems are below a *selection threshold* (dashed line in (iii)). At this point the gestures are *not* "active" in the standard sense of AP/TD. Further processes, such as an external go-signal or internal decision process, subsequently cause a transition such that gestural activations reach the selection threshold (A). At this point in time, the system of coupled planning oscillators begins to oscillate (t_3). All gestural systems are at the selection level at this point, but crucially, this does not entail that they exert strong forces on an intentional planning field. When the planning oscillator associated with a gesture reaches a triggering phase (short vertical lines in (iv)), the gestural system activation is boosted. These triggering/boosting events correspond to the initiation of gestural activation in the score (i.e. activation in the standard AP sense), and are labeled by (B) and (C) in the figure. The concepts of selection and oscillation-based triggering provide an important dissociation in the selection-coordination model: gestures are retrieved from memory and then selected for execution before they are actually executed; the precise timing of execution (i.e. gestural initiation) is governed by the system of coupled planning oscillators. A key difference between this conception and standard AP is that gestural systems have non-trivial dynamics and non-zero activation before they begin to induce movement.

Gestural force gating. In the intentional planning model, the force exerted by a gesture on a TV system is a nonlinear function of its activation. Specifically, the gain of the force is modulated by a sigmoidal gating function (vii). When the gestural force on an intentional field is strongly gated, only gestural activation values sufficiently greater than the selection level value will exert strong forces on the field. However, when the gating function is leaky, even gestural systems which are below the selection level can have detectable effects. In the example, the |LA clo| gesture is leakily gated; this entails that at timepoints t_3 and t_4, even before the gesture has been initiated, there is a non-negligible influence of |LA clo| on the LA planning field. This influence is indicated by (E) in the graph of field activity (viii) at timepoint (t_3), and its effect on the centroid of the LA field is indicated by (D) in the heat map of field activation (v). In contrast, for illustrative purposes the |ALV clo| gesture which exerts force on the TTCD field is strongly gated. Thus there is no effect of |ALV clo| on TTCD until the gesture is initiated.

The genesis of anticipatory effects. The intentional-planning model allows for anticipatory effects of gestures to occur, before those gestures are active in the standard sense, because gestural systems have subthreshold activation values before canonical activation (i.e. before gestural initiation). Thus the gestural activation intervals of a standard AP/TD gestural score are re-interpreted as periods of time in which a gestural system exerts a *relatively strong* force on a tract variable system. Why distinguish between epochs of time in which a gesture exerts a strong vs. weak force on an intentional planning field? Without such a distinction, there is no way to describe or model the control of relative timing of movement initiation, which involves relatively strong forces associated with the regime of activation induced by the selection and triggering of gestures. Such control must be independent from the relatively weak forces that result from leaky gating of gestures with subthreshold activation.

There are many questions that arise regarding the subthreshold effects predicted by the model: what determines their magnitudes? Are they language-specific, lexical-item specific, speaker-specific? How context dependent are they? Do anticipatory effects differ from perseveratory ones? Etc. The model does not preclude any particular factors from influencing the effect magnitudes, and such questions must ultimately be resolved through empirical work. Despite our uncertainty regarding the effect magnitudes, it is clear that the standard AP/TD model is unable to generate any effects of this sort whatsoever, because it does not have a concept of gestural selection. Ultimately, this concept is needed to dissociate the control of timing of gestural initiation (which involves relatively large movements and large gestural forces) from the ongoing control of vocal tract state, which may reflect relatively small influences of gestures which are part of a plan for previous or upcoming movements.

1.4 Methodological background: information and signal chopping

The speech system is unfathomably high-dimensional. We take measurements of the system—*signals*—which are comprehensibly high-dimensional, but our theoretical models posit a low-dimensional set of categories. The analytical problem we face, schematized in Fig. 3, is to quantify the relation between the signals and the theoretical categories. From an information theoretic standpoint, the appropriate quantity is mutual information, which is the information produced by observing a signal and a category together subtracted from the sum of the information produced by observing only a signal and the information produced by observing only a category. The mutual information can be visualized as the overlapping region of the circles Fig. 3, where the circles represent information associated with the signals, I(X), and with the categories, I(Y). The reader should note that these circles under-represent the difference in scale: there is a vastly larger amount of information associated with the signals than with the categories.

Although the information produced by observing a category is easy to calculate, the information produced by observing the signal is hard to calculate (for reasons we discuss below), and thus calculation of mutual information is not readily tractable. Instead of trying to calculate mutual information, this paper takes a more indirect route and uses machine learning (or more specifically, trains deep neural networks) to determine how accurately a set of signals can be used to classify members of a set of categories. The accuracy of the trained model can be taken as an indirect measure of mutual information, and we refer to it in this context as a measure of *category-related information*. Metaphorically, we can say that when a trained model exhibits above-chance classification accuracy on test data, there exists category-related information "in" the signal. This language must be viewed as metaphoric because signals are not containers and therefore it is not literally the case that anything is "in" or "inside" them. Note that the same metaphor (the container schema) is used when we say that there is "meaning in words". Additional aspects of Fig. 3 are explicated below.

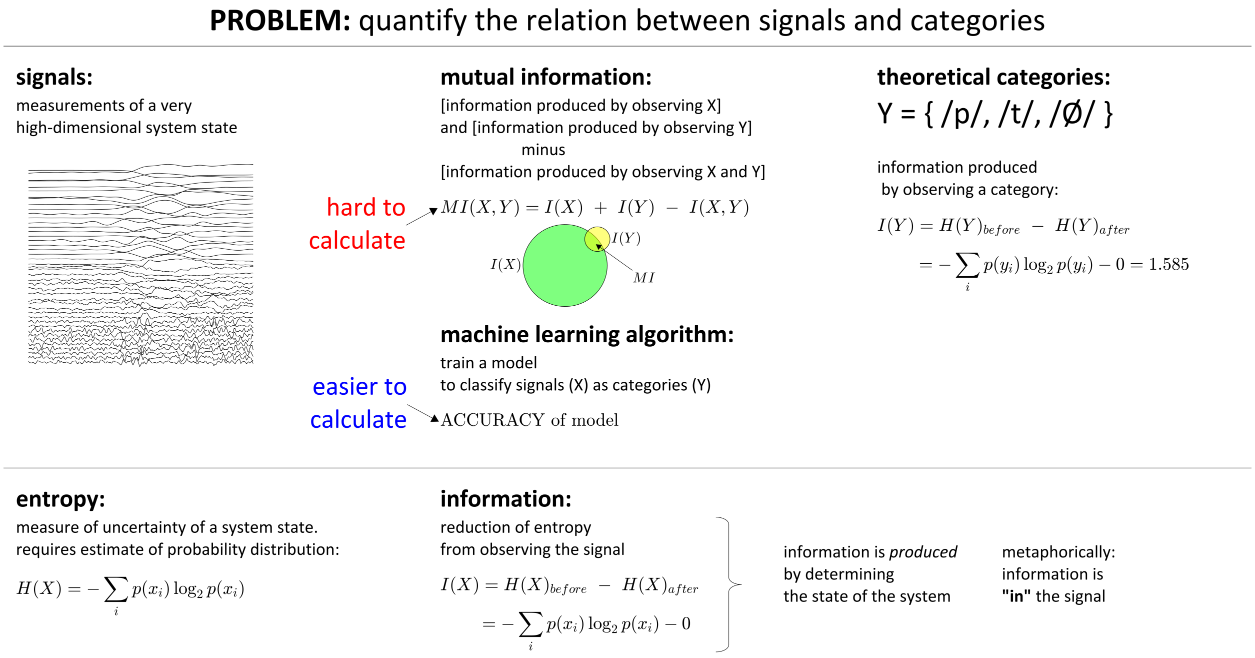

PROBLEM: quantify the relation between signals and categories
signals:
measurements of a very high-dimensional system state
mutual information:
[information produced by observing X]
and [information produced by observing Y]
minus
[information produced by observing X and Y]
hard to calculate
$MI(X,Y) = I(X) + I(Y) - I(X,Y)$
$I(X)$
$I(Y)$
MI
machine learning algorithm:
train a model
to classify signals (X) as categories (Y)
easier to calculate
ACCURACY of model
theoretical categories:
Y = { /p/, /t/, /Ø/ }
information produced by observing a category:
$I(Y) = H(Y)_{before} - H(Y)_{after}$
$= -\sum_i p(y_i) \log_2 p(y_i) - 0 = 1.585$
entropy:
measure of uncertainty of a system state.
requires estimate of probability distribution:
$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$
information:
reduction of entropy
from observing the signal
$I(X) = H(X)_{before} - H(X)_{after}$
$= -\sum_i p(x_i) \log_2 p(x_i) - 0$
information is *produced* by determining the state of the system
metaphorically: information is **"in"** the signal

Fig. 3. Overview of the problem from an information-theoretic standpoint. Signals are measurements of a very high-dimensional system state. To quantify the relation between signals and theoretical categories, information theory would use mutual information, but this is difficult to calculate. The indirect alternative we use is the accuracy of a classifier model.

To investigate category-related information in speech signals, this paper develops a novel approach to the analysis of articulatory and acoustic data: *signal chopping*. The goal of signal chopping is to characterize the spatial and temporal distributions of category-related information in a signal. In a sense, the method indirectly assesses how information associated with physical measurements (i.e. articulator positions and acoustic signals) provides information associated with presupposed categories (i.e. labial and alveolar stops). To understand this goal, it is helpful to clarify what we mean by terms such as *signal*, *information*, *category-related,* and *spatial and temporal distribution*.

Fig. 4A below shows a signal for a single production of the word *pop*. The signal here is a set of measurements that represent the state of a system, or the states of multiple systems; each line corresponds to a separate dimension of the signal. Some dimensions more directly represent the state of the vocal tract (the ARTIC dimensions, which are horizontal and vertical positions of articulograph sensors, as well as their first-differences, ΔARTIC). Other dimensions more directly represent the state of the acoustic excitation generated by a vocal tract (the MFCC and ΔMFCC dimensions).

Of course, all of these articulatory and acoustic measurements are somewhat removed from the states of the relevant physical systems, and both are necessarily drastic reductions in dimensionality. Specifically, a very high-dimensional characterization of the state of the vocal tract might include a spatially precise, 3-dimensional description of the position and velocity of all tissue-air boundaries, the tension and derivative of tension in all relevant muscle fibers, air pressure and flow at all points in space, and potentially many other state variables, including ones associated with the state of the nervous system. In contrast, the relatively lower-dimensional descriptions provided by the articulatory signals in Fig. 4 merely show the velocities and relative positions of articulograph sensors which are adhered to a few points on the lips, tongue, and jaw. Although the articulatory signals are vast simplifications of the state of the vocal tract, they are nonetheless related to its high-dimensional state. Similarly, the MFCC coefficients are derived from a somewhat complicated dimensionality reduction which indirectly represents a short-time analysis of the state of acoustic signal recorded at a microphone. These coefficients are nonetheless related to distributions of acoustic energy.

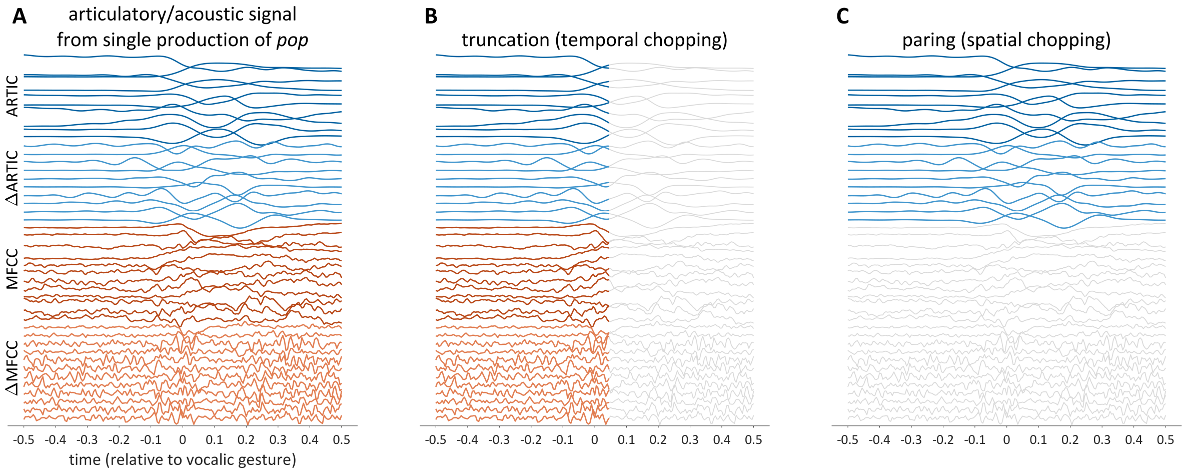


Fig. 4. Example of signal chopping for a single production of the word *pop*. (A) unchopped signal. (B) truncation, i.e. chopping in time. (C) paring, i.e. chopping in "space". The ARTIC dimensions are the horizontal and vertical positions of EMA sensors, the MFCC dimensions are Mel-frequency cepstral coefficients, and the corresponding ΔARTIC and ΔMFCC are first differences.

Each measurement, i.e. each sensor position, sensor velocity, MFCC coefficient, and MFCC coefficient derivative can be conceptualized as a separate "spatial" dimension of the system. The term *spatial* here is primarily metaphoric but one should recognize that in a physical sense the vocal tract is an entity that exists in space and that acoustic energy is a spatial pattern of energy/mass density.

Given the above understanding of a signal as a measurement which indirectly represents the state of a physical system, we can construct an understanding of what it means for *information* to be "in" a signal. Consider that every measurement in the multi-dimensional, heterogeneous signal of Fig. 4A has finite limits: there are minimum and maximum values, and a space of possible values in between. The limits are due primarily to anatomical constraints, and ultimately to the law of conservation of energy. When we take repeated measurements of the signal, we obtain a distribution of values in a space, and we can characterize this distribution in the form of a joint probability density. The entropy *H* of a distribution, which can be viewed as a measure of its degree of disorder, is defined (in bits) as $\mathrm{H} = -\sum_i p_i \log_2(p_i)$. According to standard information theory (Sethna, 2006; Shannon, 1948), the information that is *produced* by making a measurement is defined as the reduction in entropy that occurs when the observation is made. When the measurement fully determines the state of the signal, the information produced is equivalent to the entropy. To take a specific example, lets compare the entropies of a fair coin with a 0.5 probability of heads and a biased coin with a 0.8 probability of heads:

$$\mathrm{H}_{\mathrm{fair}} = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$
$$\mathrm{H}_{\mathrm{biased}} = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) \approx 0.72.$$

Before observing a coin toss, there is more uncertainty regarding the outcome for the fair coin (1 bit) than for the biased coin (≈ 0.72 bits); however, before the toss, there is no "information" in the theoretical sense—information is produced when the toss is observed and it is equal to the entropy before the toss (the entropy associated with the probability distribution over states) minus the entropy after the toss (0, since the observed state has probability 1). It is crucial to understand that information is always associated with a resolution of uncertainty. It is thus merely a metaphor to say that there is information "in" a signal; what is more accurate to say in a technical sense is that each measurement or observation of a signal resolves some amount of uncertainty and thereby produces information.

This technical understanding leads us to the next clarification, involving the phrase *category-related* information. Consider that for the words *pop* and *pot*, the coda consonants /p/ and /t/ are generally understood as different categories. For current purposes we can remain agnostic regarding the epistemological status of the categories, and we need not resolve exactly how the categories should be formally represented—perhaps they are phonemes, or segments, or perhaps the difference is a place feature. From an AP/TD perspective the word forms differ by their gestural composition; more neutrally, they are simply different motor behaviors with different acoustic consequences. What matters here is that, for a given set of signals, we have knowledge of which signals are associated with which categories. Given this knowledge, we can ask two questions: (i) which spatial dimensions of the signal contain information that can be used to predict the categories associated with the signal? and (ii) when in time is this information present? The goal of signal chopping is to answer these questions, i.e. to localize category-related information in space and in time. Since "information" is understood as a reduction of entropy, we can alternatively describe the method as determining the extent to which the reduction in entropy associated with physical measurements can be used to reduce entropy associated with observing categories.

It should be noted that the phrase "category-related information" is also appropriately described as "category-set-related" or perhaps as "category-discriminating information": it is information in signals which allows for the members of a set of categories to discriminated. Such information distinguishes between the members of a set of categories, rather than being related specifically to a particular category of the set. Categories as theoretical constructs are only sensible if we assume system of oppositions between categories, and the current analysis presupposes this.

Localization of category-related information is not an easy task, and it may not be possible to accomplish it with conventional approaches which are suited for low-dimensional data. Indeed, localization of category-related information in speech has not been attempted with conventional approaches, to my knowledge, possibly for reasons that are detailed below. In a typical, low-dimensional analysis, we might take repeated measurements of signals associated with different categories, select values from specific spatial dimensions of those signals which we believe to be relevant, and then determine when in time the distributions of values are distinct using some statistical criteria. Using *pop* and *pot* as examples, lets say we take many measurements of time-series of articulatory and acoustic signals associated with productions of these word forms. Furthermore, we believe that the aperture between the lips and the degree of constriction between the tongue tip and palate are controlled in association with /p/ and /t/, so it makes sense to use lip aperture and tongue tip position as the relevant signals. After aligning these signals over trials, we then conduct statistical analyses, at each time step, to determine whether the distributions of values from each category are likely to have been sampled from distributions with different means.

There are many problems with the above approach. First, the distributions at a given point in time might differ in their variances rather than their means; or, those distributions may be non-normal, perhaps multimodal. Second, there will be temporal and spatial correlations between the selected measurements. Third, we have ignored many other dimensions of the signal which could plausibly provide relevant information: e.g. the positions and velocities of other sensors (e.g. a sensor on the jaw). Fourth, in the acoustic domain there are no principled reasons for deciding how to reduce dimensionality: many experimenter degrees of freedom are present in the calculation of MFCCs or other time-frequency analyses. Fifth, if we conduct the analysis on many signal dimensions, separately, we would need to greatly raise our statistical significance criterion to avoid Type I error. Moreover, although adjusting for multiple independent comparisons is relatively straightforward, it is less clear how to adjust for multiple comparisons when measures are interdependent and correlated over time. Sixth, if we have more categories we will need to make more comparisons and hence more adjustments to our significance criterion.

Finally, and most problematically, just the fact that the means of two distributions do not differ statistically does not entail that the signals from each category provide “no information” regarding the likelihood of their category association—it merely entails that the parameters of underlying distributions are unlikely to differ. In fact, except in simple cases it is not straightforward to characterize the amount of category-related information which a signal contains. It is possible in principle to calculate the information in the signal, but this requires that we have accurate estimate the distributions. If those distributions are very high-dimensional, the estimates will be error-prone. Not only that, but we need to make decisions regarding how to scale each dimension—either the state spaces of values for each dimension are equally large, or they are weighted in some way. The values of mutual information measures that we obtain from any such procedure may depend intimately on such decisions. For all of the above reasons, an alternative approach is desirable.

The alternative pursued here, signal chopping, is a general procedure in which analyses of some sort are repeated on a set of multidimensional time series, while each series in the set is systematically truncated in time and/or pared in space. Examples of truncation and paring for a single production were shown in Fig. 4B and C, respectively. Each truncated/pared dataset is called a *chopped dataset* and typically includes many experimental trials, all of which are identically truncated/pared. There are no restrictions on what types of analyses can be performed on the chopped datasets, but in general, the goal of signal chopping is to characterize the spatial and temporal distribution of category-related information in a set of signals. It is important to note that the analyses performed on chopped data are conceptually and practically distinct from the chopping procedure itself. Because of the high-dimensionality of speech data (and considering the issues discussed above), a machine learning method is appropriate.

In the current study the specific machine learning approach is to train a deep neural network—consisting of bidirectional long short-term memory (biLSTM) units—to classify trials and then assess its accuracy on test data. This training/testing procedure is repeated multiple times on each chopped dataset, with random shuffling of training and test subsets. An overview of the procedure is provided in Fig. 5. The analysis is conducted separately for each subject and phonological environment (details in Section 2.4). The following steps are involved: (1) An initial dataset is constructed by extracting and aligning articulatory and acoustic signals from a number of trials; in this study the trials always belonged to one of three categories (/p/, /t/, or Ø). (2) A chopped dataset is constructed by truncating and/or paring the data. (3) Equal numbers of trials from each category are randomly assigned to training and test datasets. (4) A deep network with randomly initialized weights is trained on the training dataset, and (5) the classification accuracy of the network on the test dataset is recorded. Steps (3)-(5) are repeated multiple times for each chopped dataset. The mean classification accuracy for a given chopped dataset is viewed as an indirect measure of the extent to which information in the signal is associated with information regarding phonological categories. This allows one to characterize (albeit indirectly) how much category-related information is present in some period of time for some (sub)set of spatial dimensions of the signal; a simulated result is shown in Fig. 5.

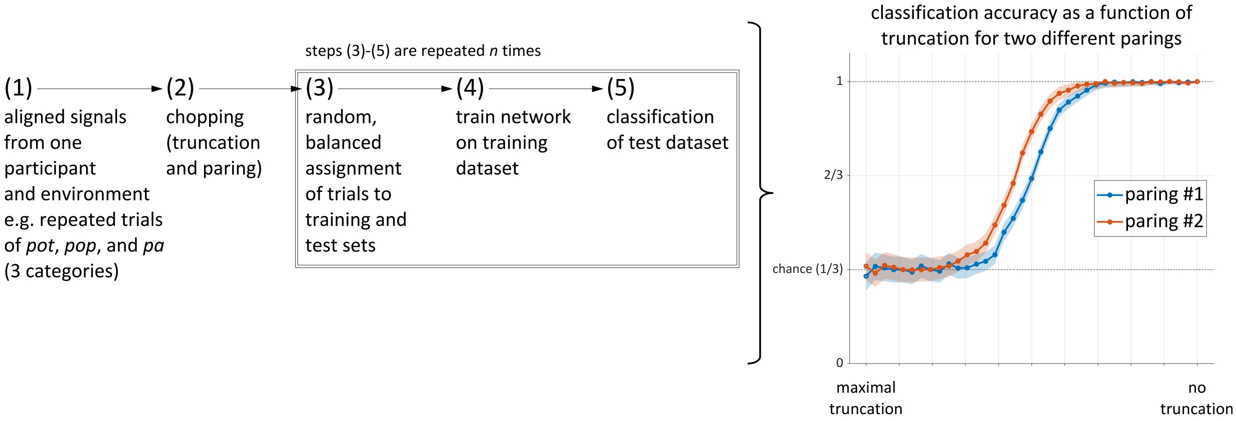


Fig. 5. Network classification analysis procedure. The graph shows simulated accuracy as a function of truncation for two differently pared datasets. Because there are three equally frequent categories in the test set, chance accuracy is 1/3 (33%).

To establish a theoretical expectation for classification accuracy—in particular one which is based on AP/TD gestural scores—it is helpful to consider how an ad-hoc optimal decision rule performs on the low-dimensional, idealized representation of speech provided by a gestural score. Consider the three instances of hypothetical 2-dimensional signals in Fig. 6A. These signals are simply continuous representations of gestural scores from the words *pah*, *tah*, and *ah*, aligned at some reference time associated with the vowel (not shown). The /t/ in *tah* involves an |ALV clo| gesture and the /p/ in *pah* involves a |LAB clo| gesture. We can interpret the states of these gestures as distinct spatial dimensions of a signal, and treat the state of each gesture as either 0 (inactive) or 1 (active). For the token of *ah* which lacks an onset gesture (indicated by /Ø/), the signal has a value of 0 in both dimensions at all times. For the other two tokens, the value increases from 0 to 1 when the consonantal constriction gesture becomes active.

One way to optimize our total classification accuracy for the signals in Fig. 6A is to adopt the following decision rule: classify a signal as /t/ if any value of the first signal dimension is 1; classify as /p/ if any value of the second signal dimension is 1; classify as Ø if no signal values in either dimension are 1. As shown in Fig. 6B, this strategy obtains 100% accuracy as long as the signals are truncated after the last gestural initiation (i.e. -50 ms, which is the initiation time for /p/). When the signals are truncated before the first gestural initiation (i.e. -100 ms for /t/), the combined accuracy is at chance (i.e. 1/3). Between initiations of /t/ and /p/ (i.e. between -100 and -50 ms), the optimal accuracy is 67%. At all times, the total predicted accuracy is the average of the within-class accuracies.

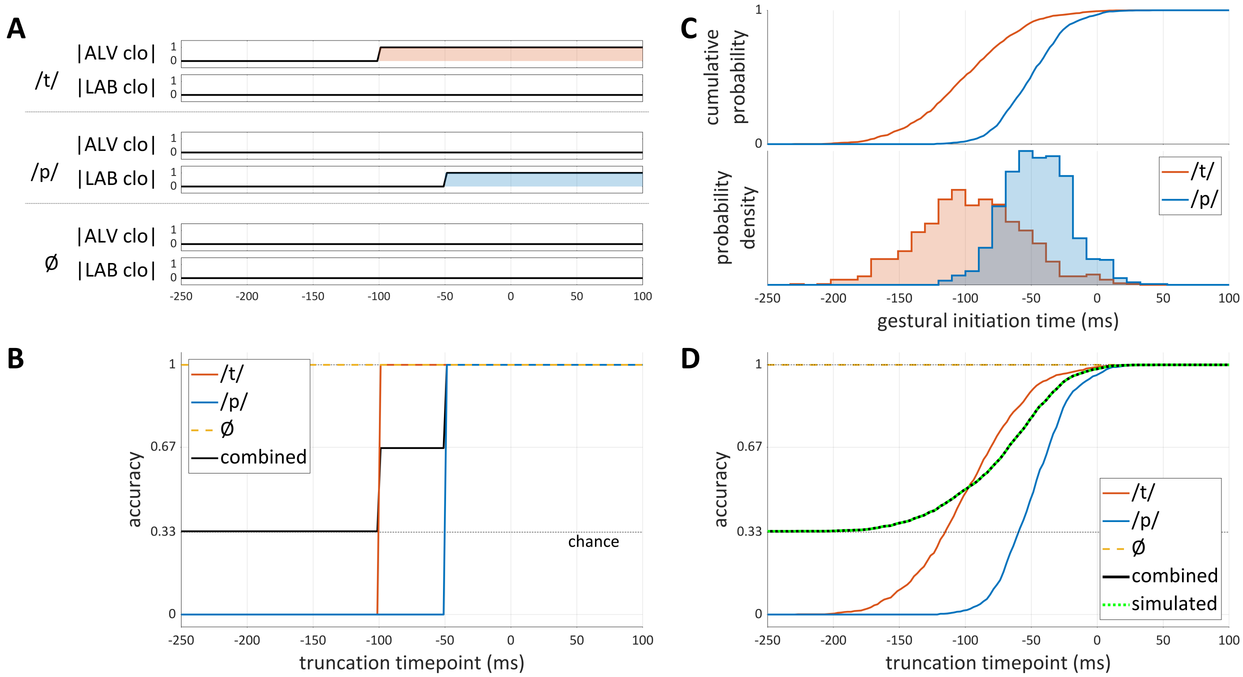


Fig. 6. Ideal classification accuracy based on gestural score representations. (A) Gestural scores for /ta/, /pa/, and /a/ as two-dimensional signals. (B) Within-class and combined accuracy as a function of signal truncation, using an optimal decision rule. (C) Cumulative distribution and probability density of simulated gestural initiation times. (D) Within-class and combined accuracy, using an optimal decision rule. A transformation of the cumulative distribution functions of simulated data (green dashed line) is equivalent to combined accuracy.

When we scale up the analysis of ideal classification from 1 token of each class to 1000 tokens of each class, we can see that the within-class accuracies are directly related to the cumulative distribution functions of the times at which the relevant gestures are initiated. Probability densities and cumulative probabilities of gestural initiation times obtained from random sampling of Gaussian distributions (i.e. simulated experimental data) are shown in Fig. 6C. Using the same decision rule as before and chopping the data with truncation, we obtain 33% (chance) accuracy for truncations before any gestures are initiated, and there is a sigmoidal increase in accuracy which reaches 100% for truncations after all gestures have been initiated (Fig. 6D). Indeed, the optimal classification accuracy is equivalent to (1/3)(1 + *cdf*[p] + *cdf*[t]), labeled as "simulated" in the figure. This means that the optimal classification accuracy based on a set of gestural scores can be calculated from a transformation of the cumulative distribution functions of empirical measures of when gestures are initiated. The importance of this is that we can easily derive a predicted classification accuracy based on empirical estimates of when gestures are initiated, and this optimal prediction provides a comparison that is useful for interpreting the network-based analyses. Note that treating the Ø category as a third dimension of a score with a positive feature value does not change the general result: in that case the optimal rule would be to guess randomly unless there is non-zero value in any dimension, and the average accuracy has same sigmoidal shape as in Fig. 6D.

1.5 Hypotheses

Below we delineate the hypotheses of the current study. Articulographic and acoustic data were collected for (C)V(C) forms for all combinations of /p/, /t/, and /Ø/ onsets and codas, with the intervening vowel /a/. Furthermore, all responses were preceded by a prolonged /i/. Thus, there are two main contexts in which anticipatory information can analyzed: prior to an onset and prior to a coda. For the onset context, there is one preceding environment: /i__/. For the coda context, there are three preceding environments: /ipa_/, /ita_/, and /iØa_/.

Hyp: *Pre-initiation intentional planning:* Due to subthreshold gestural influences on the state of the vocal tract, there is category-related information in speech signals prior to gestural initiation. This predicts that classification accuracies for onsets or codas will be significantly above chance before the relevant estimates of movement initiation. The corresponding null hypothesis is the standard AP/TD prediction that accuracy functions should reflect the cumulative density functions of movement initiation times. Note that gestural initiation times closely correspond to estimates of movement initiation (see section 1.1 and Appendix).

Hyp: *Onset-coda interference*: Anticipatory effects of codas will be stronger when there is no syllable onset. For example, the anticipatory effects of the coda in the form /ap/ will be stronger than in /pap/, because in /pap/ there is an active onset gesture which drives the lip aperture tract variable system. This hypothesis predicts that coda classification accuracy will rise above chance earlier in onsetless responses than in responses with /p/ or /t/ onsets.

Hyp: *Articulatory-acoustic information asymmetry*: Given that there are multiple ways to configure the articulators to achieve similar acoustic results, some articulators may be positioned in a way that anticipates an upcoming gesture without inducing any measurable change in the acoustic signal. Thus there may be more information in articulatory signals than acoustic ones. For a specific example, the jaw might be elevated to a greater degree prior to /p/ and /t/ onset responses than it is prior to /Ø/ onset responses, and yet, the preceding vowel /i/ might not be acoustically distinct between these conditions because the tongue body and lips can be lowered or raised to compensate for differences in jaw elevation. This hypothesis predicts that classification accuracies should be higher for networks trained on only articulatory data than for networks trained only on acoustic data.

Hyp: *Articulator-specific information*: Some articulators may be substantially more informative than others regarding category identity. This very general hypothesis predicts that classification accuracies will be differently affected by paring data to exclude articulatory signals from one sensor at a time. No specific predictions are held regarding these differences; instead the aim is to investigate the relative contributions of articulators in an exploratory fashion. Note that informativity of particular signal dimension in this context relates to its utility in discriminating all of the categories, rather than its association with any particular category.

2. Method

2.1 Participants and task

Six native speakers of English participated in the study. During the experiment participants produced nine different target forms. The target forms were all possible CVC syllables with /p/, /t/, and Ø onsets and codas, and the vowel /a/. The visual stimuli that cued the target forms were the orthographic representations shown in Table 2. For two stimuli, *ot* and *op*, the experimenter demonstrated the desired pronunciations (i.e. [at] and [ap]) when giving instructions. No participants had difficulty with producing the correct forms, and in all cases participants interpreted the vowel of the target form as the low central vowel /a/.

Table 2. Target stimuli

		coda		
		p	t	Ø
	p	*pop*	*pot*	*pah*
onset	t	*top*	*tot*	*tah*
	Ø	*op*	*ot*	*ah*

Participants were seated in front of a computer monitor during the experiment. At the beginning of each trial, the target stimulus appeared in the center of the screen and three dots appeared below it. The dots disappeared one by one at 500 ms intervals. During this time, participants produced a prolonged /i/ vowel, which we will call the *pre-response vowel*. 500 ms after the last dot disappeared, a green box appeared in the background of the target stimulus. The green box served as a go-cue which signaled participants to produce the target word. 1500 ms after the go-cue, the stimulus text and background disappeared. Recording continued for an additional 500 ms. The full time-course of a trial is illustrated in Fig. 7.

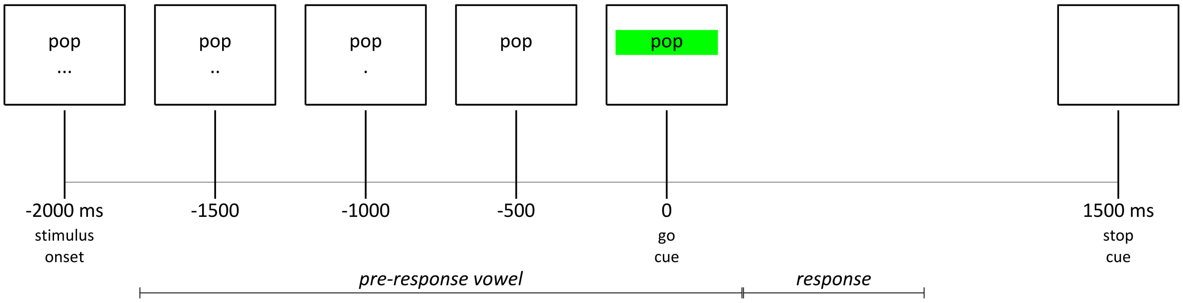


Fig. 7. Time course of a single trial. The target stimulus (*pop* in this example) appears 2000 ms before the go-cue, which is a green box in the background. During the pre-response period, the speaker produces a prolonged /i/ vowel.

The purpose of the pre-response vowel was to encourage participants to adopt a consistent vocal tract posture prior to production of the response. The vowel sequence /i_a/ was selected in order to maximize the movement range of the tongue body. The dots provided a countdown to the go-signal, in order to reduce variance in the initiation of the target. Participants were instructed to produce the pre-response vowel continuously until they began production of the target, and to begin production of the target immediately when the go-signal appears, but no earlier than the go-signal. Furthermore, participants were instructed to say the target words as consistently as possible throughout the experiment, and to avoid intentionally changing their productions.

Each participant performed 15 or 16 blocks of 36 trials during a session. In each block all nine stimuli were repeated four times in random order. A total of 540 trials (15 blocks) or 576 trials (16 blocks) were performed by all participants.

2.2 Data collection and processing

Acoustic data were collected during each trial at 22050 Hz with a shotgun microphone positioned approximately 1.5 m from the participant. Articulatory data were collected with an NDI Wave electromagnetic articulograph with a sampling rate of 400 Hz. Reference sensors used for head movement correction were located on the nasion and the left and right mastoid processes. Articulator sensors were located midsagittally on the upper lip (UL), lower lip (LL), gingiva below the lower incisors (JAW), tongue tip (TT) 1 cm from the apex of the tongue, and tongue body (TB) 5-6 cm from the apex of the tongue.

To prepare articulator data for analysis, the following procedure was applied. Reference sensor and articulator positions were low-pass filtered at 5 and 10 Hz, respectively, using 4th order Butterworth filters. After correction for head movement, the horizontal and vertical coordinates of articulator sensors were upsampled to 1000 Hz, which allows for more precise estimation of kinematic landmarks. A lip aperture (LA) time series was defined by calculating the Euclidean distance between the UL and LL sensors, and RMS velocities of all sensors were calculated.

For network-based analyses, acoustic data from each trial were transformed to MFCC matrices using a 30 ms window and 5 ms time steps. A third-party Matlab toolbox (Kamil Wojcicki, HTK MFCC Matlab) was used to calculate MFCCs, with the following parameter values: frequency range: 300-5000 Hz; number of cepstral coefficients: 12; preemphasis factor: 0.97; number of filterbank channels: 20; cepstral sine lifter parameter: 22.

Acoustic durations of segments were obtained via forced alignment using Kaldi (Povey et al., 2011). For each speaker, one token of each target was randomly selected and all segments in the response were manually labeled. For onsetless syllables (i.e. /a/, /ap/, /at/), an onset segment was included because in the transition from the pre-response vowel [i] to [a], speakers produced either a full glottal stop or exhibited a period of irregular vocal fold vibration, suggestive of a glottal constriction gesture. Monophone models were initialized and trained on the manually labeled data from all speakers. These were used for subsequent acoustic alignment of all trials.

For each participant, all data from the first block of trials (i.e. 36 trials) were excluded from analyses because these trials tend to be more variable: participants usually need several trials to become familiar with the pacing of the task and the stimuli, and they become more accustomed to speaking with the articulograph sensors over this period of time. Malfunctions in the collection of audio data necessitated the exclusion of 80 out of 3204 trials from the dataset (2.5%); almost all of these were from one participant (P5, 76 trials). A handful of trials on which participants made an error (0.13%) were excluded as well.

2.3 Articulatory landmark identification

Articulatory kinematic landmarks were obtained as follows, beginning with a global alignment procedure which aligns the articulatory signals from trials of each target, for each participant. Note that the global alignment is used solely to facilitate kinematic landmark identification, and does not directly determine the input to the signal chopping procedure. First, all trials for a given target response and participant were globally aligned using the RMS velocities of LL, JAW, TT, and TB sensors, extracted from 200 ms before the go-signal to 2000 ms after the go-signal. The RMS velocities were smoothed with a 50 ms moving average filter for this purpose. Fig. 8A shows LL and TB RMS timeseries for all trials of /pat/ from participant P01, before the global alignment procedure. Fig. 8B shows the same trials after the global alignment. The global alignment procedure was conducted as follows. (1) Calculate the average RMS timeseries for each sensor *j* over trials (μ_j^{RMS}). (2) For each RMS timeseries from trial *i* and sensor *j* (i.e. RMS_{ij}), calculate the sample lag of maximal cross-correlation ($\tau_{i,j}^{max}$) between RMS_{ij} and μ_j^{RMS}. (3) Calculate the average of $\tau_{i,j}^{max}$ over sensors for each trial ($\hat{\tau}_i$). (4) Shift all RMS_{ij} by $\hat{\tau}_i$. The procedure is iterated (and μ_j^{RMS} is updated in each iteration) until the sum of absolute values of $\hat{\tau}_i$ over trials is less than or equal to one sample. Because the period of time in which articulators are moving rapidly is relatively short, and because the movement of the tongue body (TB) from the pre-response [i] posture to an [a] posture tends to have the largest RMS, the procedure is very effective for obtaining an alignment of trials in which the TB RMS velocity maxima are approximately aligned (see Fig. 8B, red line). On the basis of this approximate alignment, the precise locations of the TB RMS velocity maxima from each trial can be identified, as well as RMS velocity maxima associated with constriction gesture initiations (i.e. LA RMS maxima for /p/, and TT RMS maxima for /t/).

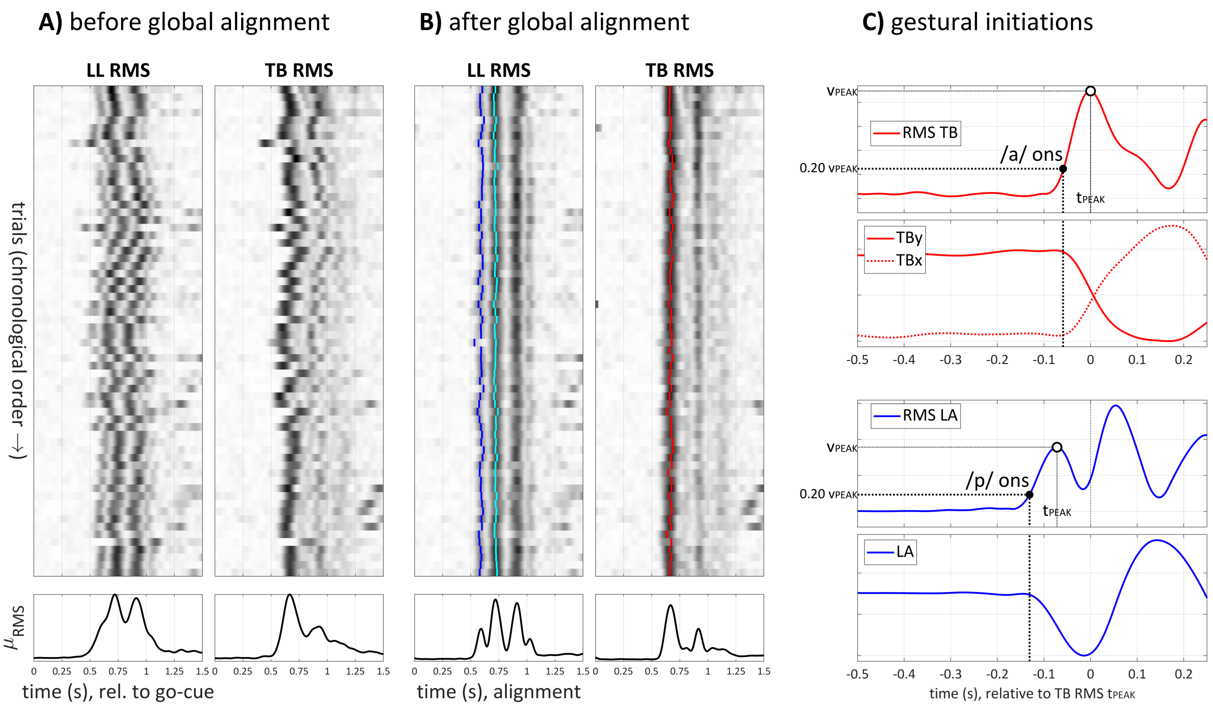


Fig. 8. Illustration of landmarking procedures. (A) before alignment; (B) after alignment. (A) and (B) show LL RMS and TB RMS for /pat/ trials from subject P01, along with μ_{RMS}. (C) Single-trial example of estimation of vowel gesture and onset consonant gesture initiations, using TB RMS and LA RMS, respectively. The vertical lines (gestural initiation estimates) correspond closely to the movement onsets evident in the LA and TBx/TBy panels.

Subsequent to the global alignment and location of RMS velocity maxima, onset and coda gestural initiations were identified as follows. For each trial and onset/coda gesture, the relevant sensor RMS timeseries (LA for /p/, TT for /t/) was smoothed with a 10 ms rectangular filter, and the RMS maximum associated with a constriction was selected as the most temporally proximal peak to the one identified in the global alignment. The RMS peak velocity value v_{peak} and time t_{peak} are shown in the examples in Fig. 8C. The gestural initiation time is estimated as the last time before t_{peak} at which the RMS velocity is less than 0.20 × v_{peak} (i.e. 20% of the maximal RMS velocity). As can be seen in the TBy, TBx, and LA time series in Fig.

8C, this procedure results in fairly accurate estimates of when the relevant tract variables begin to change substantially.

2.4 Signal chopping and network analysis procedure

Signal chopping is a general procedure in which analyses of some sort are repeated on a set of time series (see also Section 1.4), with systematic manipulation of the temporal window (i.e. truncation) and spatial dimensions of the signals (paring). A maximal dataset is defined as an unpared, untruncated set of data. In the current analyses, the maximal datasets included 20 articulatory dimensions, which were the horizontal and vertical positions and velocities of UL, LL, JAW, TT, and TB sensors, and 24 acoustic dimensions, which were the 12 MFCC coefficients and their first-differences. Both acoustic and articulatory signals were sampled at 200 Hz, i.e. steps of 5 ms. All datasets were aligned to the time of maximal TB RMS velocity associated with the retraction of the tongue root and lowering of the tongue body in the [i]-to-[a] transition (see Section 2.3). This event was used for alignment because it is robustly detectable for all responses in the experiment; note that the timing of this event may be systematically affected by the onset consonant and this is taken into account in metrics of anticipation derived below.

The maximal datasets are listed in Table 3. Network analyses were always conducted on data from one participant at a time. The reason for this is that different participants will have different vocal tract dimensions, resulting in speaker-specific variation in articulatory and acoustic values. Speakers will also produce responses with different durational profiles. This cross-speaker spatial and temporal variation cannot be readily factored out of the data, and would adversely affect network classification accuracies if network training were conducted across speakers. Furthermore, the datasets used for coda classification were separated by onset identity—i.e. chopping procedures were applied separately for data from /pa_/, /ta_/, and /Øa_/ environments. The reason for this is that information associated with variation in the identity of the onset consonant interacts with information relevant to predicting the identity of the coda. Hence the network can learn to classify codas more effectively if onset-related variation is not present in the training data. Separating the coda environments in this way also allows for analysis of the influence of onset identity on coda-related anticipatory information.

Table 3. Dataset information and independent variables

dataset name	target position	has onset	phonol. environ.	dataset size[1] (trials/partic.)	maximal window[2] (ms)	minimal window[2] (ms)	num. win.
full	onset	n.a.	/i_a/	540	[-1000, 100]	[-1000, -900]	31
onset_p	coda	yes	/ipa_/	180	[-500, 500]	[-500, -400]	37
onset_t			/ita_/				
onset_Ø		no	/iØa_/				

[1] Dataset sizes are somewhat less than these values because of excluded data (see Section 2.2).

[2] Windows are specified relative to a reference event time, the maximal TB RMS velocity (see Section 2.3).

For onset classification, the maximal dataset window was from 1000 ms before to 100 ms after the reference time. Over chopping iterations, the endpoint of this window was decreased in steps of 25 ms to -400 ms, and then in steps of 50 ms to -900 ms. For coda classification, the maximal window of analysis was from -500 to +500 ms relative to the reference time. Over chopping iterations, the endpoint of this window was decreased in 25 ms steps to -400 ms.

Paring (i.e. spatial chopping) was accomplished simply by removing signal dimensions. Hence in subsequent analyses the “combined” data include both articulatory and acoustic channels, while “articulatory” and “acoustic” datasets are constructed by paring the acoustic or articulatory dimensions, respectively. Furthermore, to investigate articulator-specific information, chopped articulatory datasets were constructed by paring all position and velocity dimensions from one sensor at a time (i.e. from TB, TT, JAW, LL, or UL).

The network training procedure is as follows. For a given chopped dataset, half of the trials are randomly assigned to a training set, and the other half to a test set. The training data are then used to train a bidirectional LSTM neural network to classify the relevant segments (i.e. onsets or codas, depending on which dataset is involved). This random partitioning of the data into test and training sets was repeated 20 times for each chopped dataset. Thus the total number of networks trained for analyses presented here was 6 speakers x [(31 truncations x 8 parings x 1 context) + (37 truncations x 8 parings x 3 contexts)] x 20 repetitions = 136,320 networks.

The accuracy of the trained networks on the test set is taken as an indirect measure of the amount of information in the input signals which is relevant to the identities of the classified segments. A deep neural network with bidirectional long short-term memory units (biLSTM) was used because these networks have been shown to be highly effective in speech recognition and because they resolve some of the problems that arise with vanishing gradients in RNNs (Hochreiter & Schmidhuber, 1997). The network structures and training algorithm parameters were identical for all analyses. Details of the network structure and training procedure are provided in Appendix: Network details. Most of these details are not directly relevant to the main issues of this paper, but the reader should understand that it is not possible to optimize or motivate all of the network design hyperparameters that are involved. Because of this, the accuracy results reported here are necessarily *under*-estimates of the maximal accuracy that one could obtain with infinite computing power, infinite time, and knowledge of the optimal network structure.

To assess the presence of category-related anticipatory information, the network accuracy on test sets is analyzed as a function of truncation. Each accuracy calculation on a chopped dataset was repeated 20 times, with each repetition having a separate random assignment of trials to training and test sets and a separate network. Accuracy functions were obtained by interpolating the mean accuracies over the full range of truncations (5 ms steps, cubic interpolant). To quantify when in time there is a substantial amount of category-related information present in a signal, or how much information is present, three anticipation metrics were constructed. Examples for two speakers are shown in Fig. 9, where accuracies are plotted as a function of truncation time (i.e. the endpoint of the analysis window, defined relative to the alignment point, TB RMS velocity maximum). Note that the AP/TD-based accuracy prediction (solid red line) is defined as in Section 1.4, based upon estimated gestural initiation times and an optimal classification strategy. The score-based prediction is calculated as $(1/3)(1 + p_{cdf} + t_{cdf})$, where p_{cdf} and t_{cdf} are gestural initiation cumulative distribution functions for bilabial and alveolar closure gestures, respectively.

One of the three metrics (ΔA) has temporal units and aims to capture how early there is substantial category-related information, relative to when such information would be expected on the basis of gestural initiation. To that end, ΔA is defined as the first truncation time at which accuracy is 10% above chance (i.e. 0.10 + 0.33) minus the mean time of gestural initiation. This corresponds to the dark green lines in Fig. 9. Note that confidence intervals for accuracy functions (±2 s.e. calculated over network training/test repetitions) are typically smaller than 10%; thus a metric that corresponds to the first truncation time at which accuracy is significantly greater than chance would provide even earlier estimates of when category-related information is available. The more conservative threshold of 10% above chance is used for ΔA in order to avoid spurious estimates that might arise from Type I error. Note also that the mean time of gestural initiation is defined by-participant, by-position (i.e. onset/coda), and by-environment (i.e. pa_, ta_, and Øa_ for codas); furthermore, when ΔA is calculated within categories (i.e. separately for classification of p, t, and Ø) as in Section 3.2, the corresponding mean gestural initiations for the category are used (for Ø, which has no initiation, the average of /p/ and /t/ initiations is used).

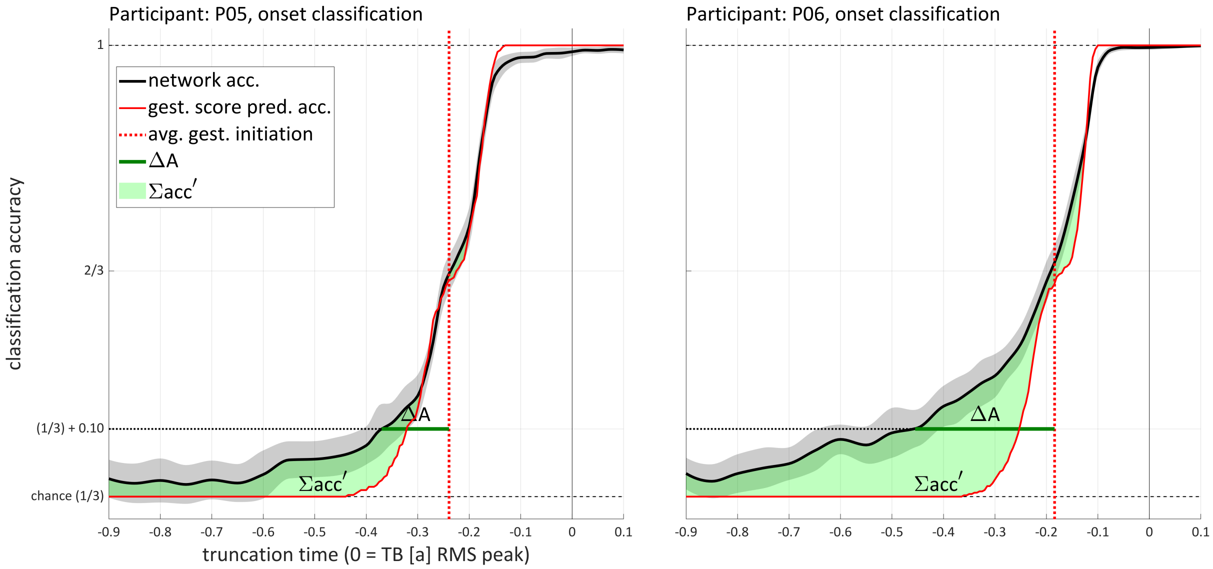


Fig. 9. Example of anticipation metrics. Network (black line) and gestural score-based (solid red line) accuracies are shown as a function of truncation time. ΔA is the time at which network accuracy first reaches a threshold of (1/3) + 0.10, i.e. 10% above chance, minus the average time of gestural initiation. Σacc′ is the area between the network accuracy curve and AP/TD-predicted accuracy where the former is greater than the latter.

The other two metrics integrate accuracy over truncation times. The metric μ_{acc} is the average network accuracy, i.e. the average value of the network accuracy function over all truncations. The metric acc′ is the excess accuracy expressed as a proportion of the maximum possible excess accuracy. The excess accuracy is the sum of the differences between the network accuracy function and predicted accuracy function, counting only positive values (i.e. Σacc′ in Fig. 9, corresponding to the area of the green region). The motivation for acc′ is that μ_{acc} does not take into account the expected accuracy, which is based on empirical gestural initiations. Only positive values are counted because in general, network accuracy is lower than expected accuracy in the temporal vicinity of gestural initiation; it is mainly in the period of time prior to gestural initiation where the network exhibits an ability to detect more category-related information than predicted by the standard AP/TD model. Note that μ_{acc} is sensitive to the range of truncations which are analyzed and hence cannot be compared between onset and coda-classification environments. μ_{acc} is also biased by cross-participant variation in gestural initiation: it is correlated with the timing of gestural initiation relative to the vocalic reference. The measure acc′ is much less sensitive to the range of truncations, and is not biased by variation in gestural initiation.

To analyze the effects of dataset characteristics and paring on anticipatory information, linear mixed effects regressions were conducted on ΔA, μ_{acc}, and acc′ with participant as a random intercept. Log-likelihood tests were used to assess the significance of regression model predictors. When significant effects were found, post-hoc t-tests were conducted.

3. Results

For all participants in all environments there was category-related information prior to gestural initiation. This contradicts the predictions of standard AP/TD and supports the hypothesis of *pre-initiation intentional planning*. Furthermore, evidence in support of the *onset-coda interference hypothesis* was obtained: category-related information regarding codas was present earlier in the /Øa_/ environment than in the /pa_/ or /ta_/ environments. The hypothesis of *articulatory-acoustic information asymmetry* was supported as well: more information was present in articulatory signals than acoustic signals. Regarding the hypothesis of articulator-specific information, only one substantial articulator-specific effect was observed.

3.1 Anticipatory information for onsets and codas

For all participants and environments, network accuracy functions obtained from combined articulatory and acoustic signals showed above-chance classification before gestural initiation. Fig. 10 and Fig. 11 show accuracy for onsets and for codas in the Øa__ environment, respectively. Accuracy is plotted as a function of truncation time, using the combined data (i.e. articulatory and acoustic signal dimensions). The black lines show the mean accuracy from network analyses, the gray bands show ±1 st. dev. calculated over repetitions, and the red lines show the AP-predicted accuracy based on gestural initiation (see Section 1.4). In almost all cases the network accuracy reaches above-chance levels before the AP prediction, but this effect is stronger for some participants than others (e.g. P01, P04, and P06 have more extensive anticipatory information than P02, P03, and P05).

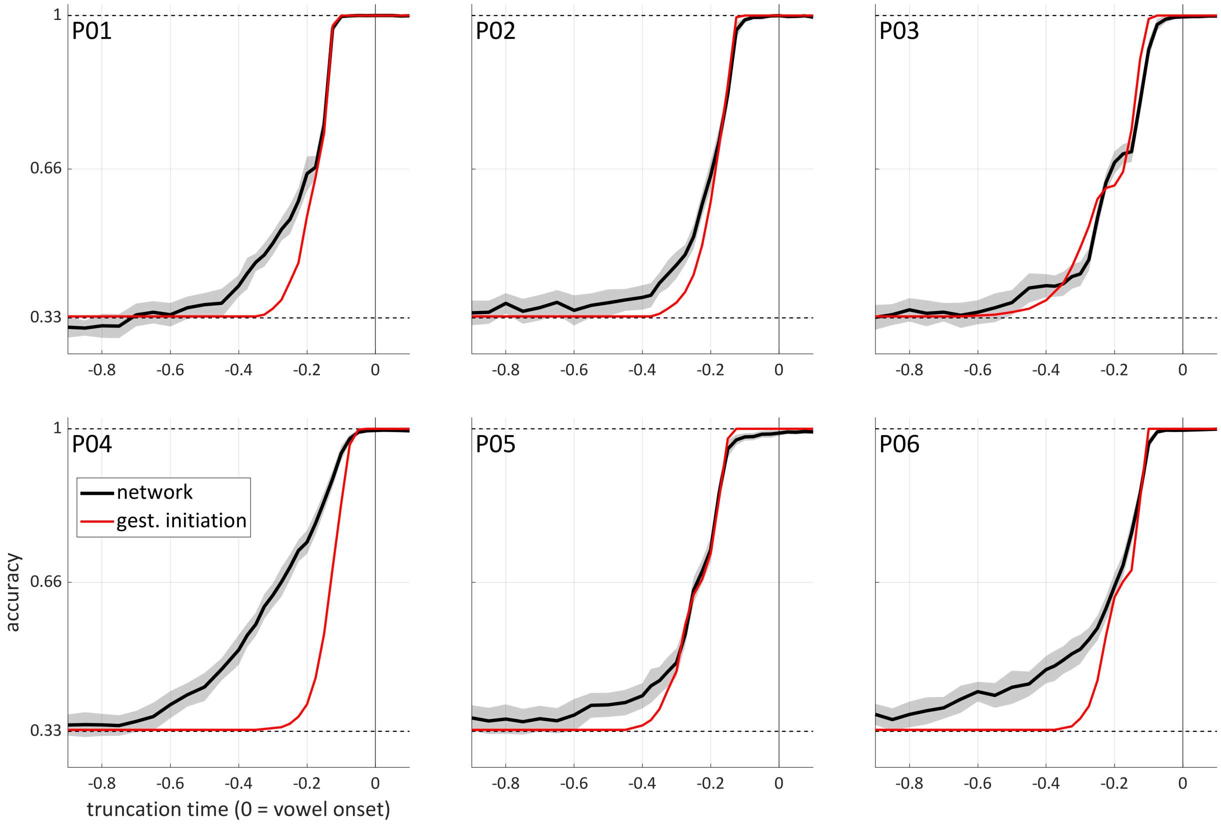


Fig. 10. Onset identification accuracies as a function of truncation. Black lines show mean accuracy, gray bands ±1.0 st. dev. Red lines show predicted accuracy based on gestural initiation.

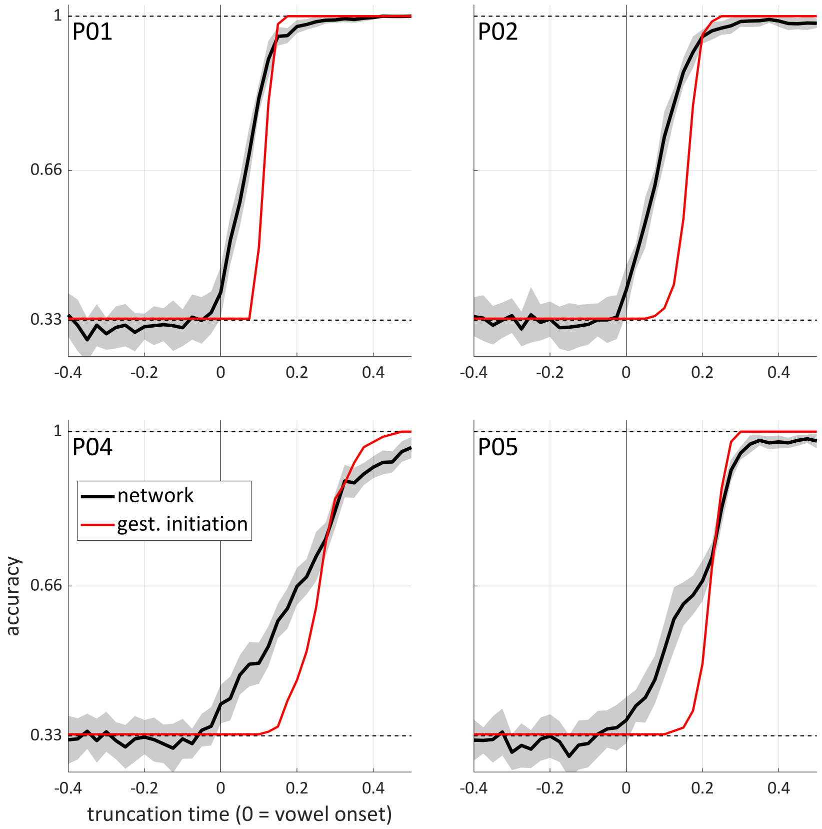


Fig. 11. Coda identification accuracies as a function of truncation, in the Øa__ environment. Black lines show mean accuracy, gray bands ±1.0 st. dev. Red lines show predicted accuracy based on gestural initiation.

The anticipation metrics ΔA, μ_{acc}, and acc′ are shown for all participants/environments in Fig. 12. As defined in Section 2.3, ΔA represents how early before gestural initiation there exists substantial category-related information. Mixed effect linear regressions of the anticipation metrics showed a significant effect of position (onset vs. coda) on ΔA ($\chi^2(1) = 6.5$, $p < 0.05$): anticipatory information tended to occur earlier for onset categories than for coda categories (ΔA: onset - coda = -79 ms). Position was not a significant predictor for μ_{acc} or acc′. The reader should note that μ_{acc} is not directly comparable between positions because the ranges of analysis windows differed.

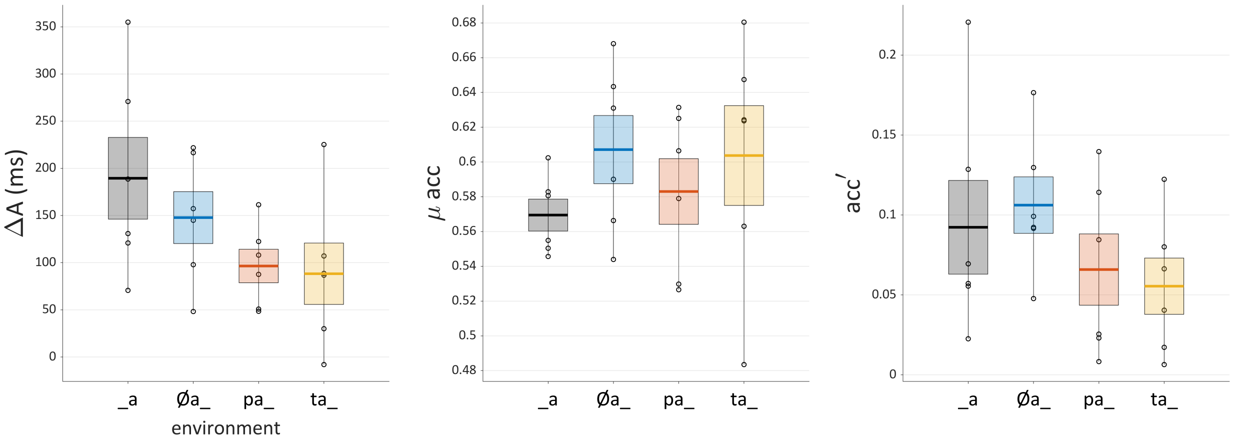


Fig. 12. Anticipation metrics for onsets (_a environment) and codas (pa_, ta_, Øa_ environments). Horizontal lines show across-participant mean accuracy, boxes show ± 1 s.e., circles show individual participant accuracy. Note that μ_{acc} is not directly comparable between onset vs. coda positions because the ranges of analysis windows differed.

Within the coda position datasets, regressions showed a significant effect of phonological environment on ΔA ($\chi^2(1) = 4.3$, $p < 0.05$): anticipatory information occurred earlier for /Øa_/ (ΔA: onsetless – with-onset = -55 ms). There was also a significant effect on acc′ ($\chi^2(1) = 9.4$, $p < 0.01$), with a greater proportion of excess anticipatory information in the /Øa_/ environment than in the /pa_/ and /ta_/ environments (acc′: onsetless – with-onset = 0.045). The greater degree of anticipatory information in the /Øa_/ environment was predicted by the *onset-coda interference hypothesis*, the rationale being that articulators recruited for a coda constriction gesture are less constrained in the pre-coda epoch when there is no active onset constriction gesture.

3.2 Segment-specific differences in anticipatory information

There were no significant segment-specific differences in classification accuracy, when variation in gestural initiation is taken into account. As can be seen in Fig. 13A (middle panel), the only significant effect of target category was found in onset position for the metric μ_{acc} ($\chi^2(2) = 7.0$, $p < 0.05$). Post-hoc paired t-tests showed that the /t/ was detected earlier than Ø and /p/ categories. However, these differences are not unexpected since /t/ gestural initiations were earlier than /p/ initiations, at least when defined relative to the reference event of TB RMS velocity maximum. When the earlier initiation time of /t/ is taken into account, which is the case for the ΔA and acc′ metrics, no significant differences are observed.

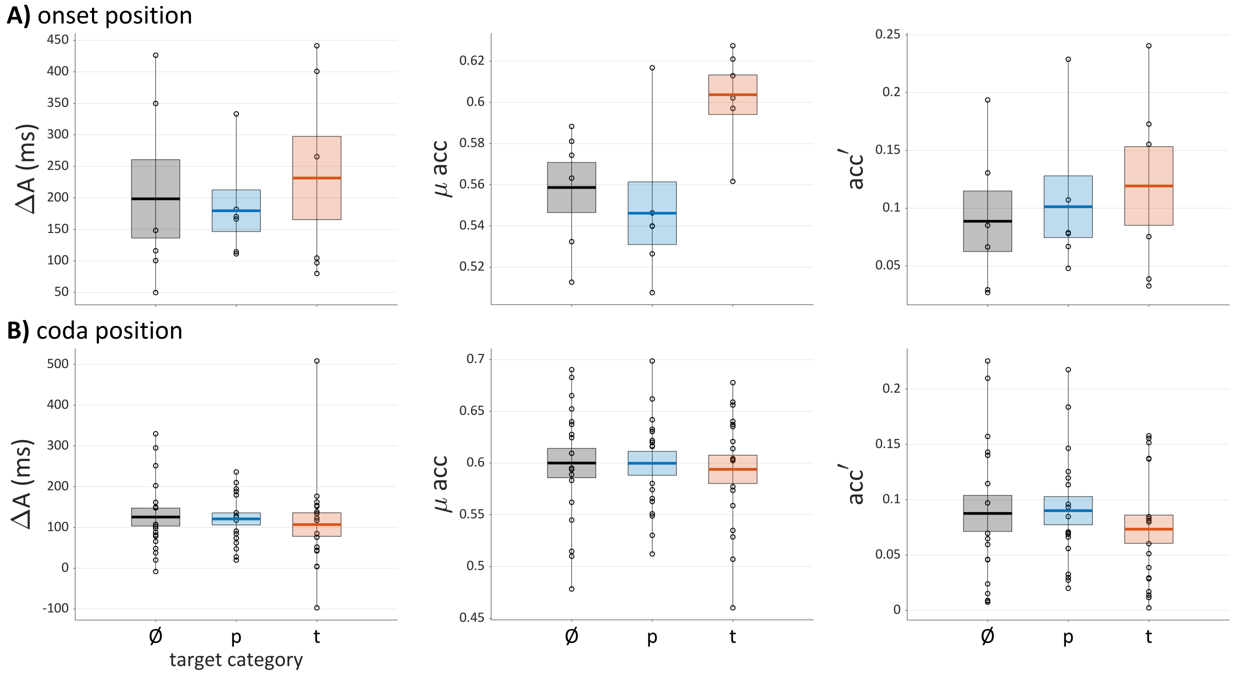


Fig. 13. Segment-specific differences in anticipation metrics. (A) onset position. (B) coda position. Horizontal lines show across-participant mean accuracy, boxes show ± 1 s.e., circles show individual participant accuracy (i.e. metrics for each speaker/target category).

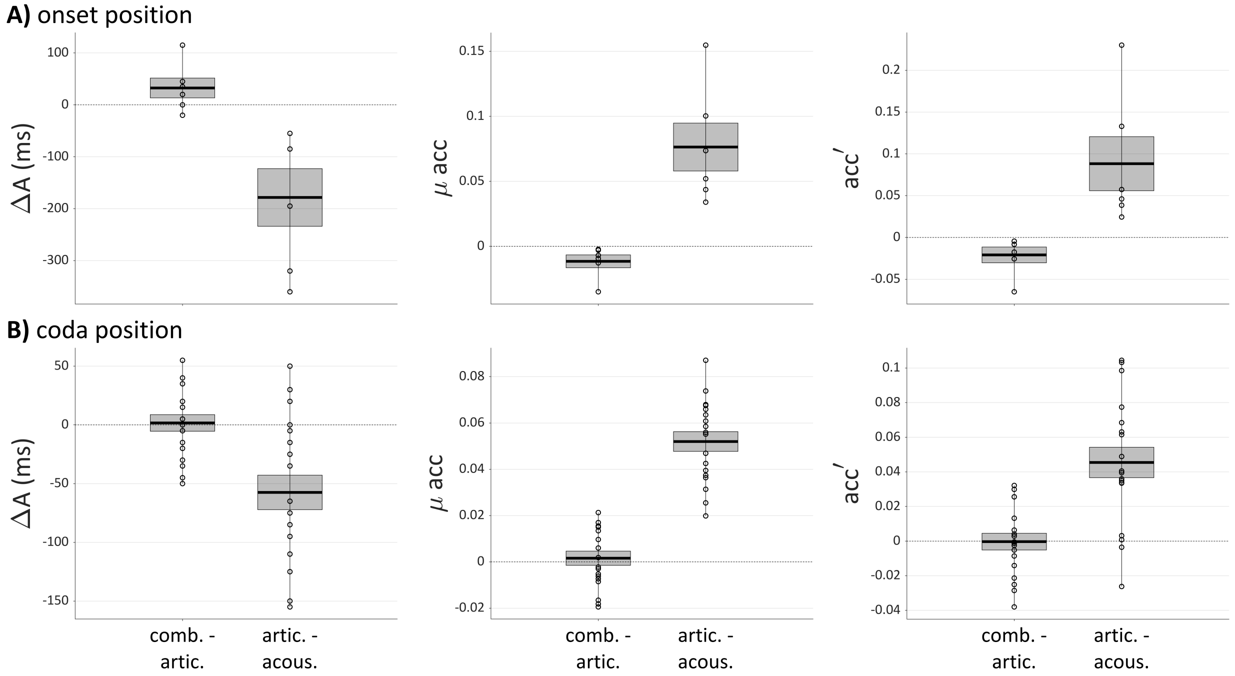
3.3 Effects of information source

Articulatory data contained more category-related information than acoustic data, but the combination of articulatory and acoustic sources did not lead to substantial improvements in accuracy; this suggests that the acoustic information is mostly redundant with articulatory information. For all three metrics, information source was a significant predictor in onset position (ΔA: $\chi^2(2) = 12.9$, $p < 0.01$; μ_{acc}: $\chi^2(2) = 18.9$, $p < 0.001$, acc′: $\chi^2 = 10.1$, $p < 0.01$) and in coda position (ΔA: $\chi^2(2) = 11.2$, $p < 0.01$; μ_{acc}: $\chi^2(2) = 44.6$, $p < 0.001$, acc′: $\chi^2 = 21.6$, $p < 0.001$). Table 3 shows results of post-hoc paired t-tests. For all six combinations of anticipation metrics and positions, there were no significant differences between the combined data (comb.) and the articulatory data (artic.), which suggests that the acoustic data (acous.) did not add information beyond what was present in the articulatory data.

In contrast, all six combinations of metric and position differed significantly between articulatory and acoustic data. The effects in ΔA indicate that category-related information was present earlier in articulatory data than in acoustic data, and the effects in μ_{acc} and acc′ indicate that there was more category-related information overall in articulatory data than acoustic data. The effects are also shown in Fig. 14 which plots the by-participant/environment differences in metrics for both onset and coda positions. It is evident that the differences between metrics for combined vs. articulatory data are close to zero, while these differences depart substantially from zero for articulatory vs. acoustic data.

Table 3. Post-hoc tests of anticipation metrics by information source

	onset position			coda position		
metric	Δ	t-stat. (df)	p-value	Δ	t-stat. (df)	p-value
ΔA	comb. - artic. = 0.032	t=1.7 (5)	p = 0.15	comb. - artic. = 0.002	t=0.2 (17)	p = 0.82
	artic. - acous. = -0.178	t=-3.2 (5)	p < 0.025 *	artic. - acous. = -0.057	t=-3.9 (17)	p < 0.025 *
μ_{acc}	comb. - artic. = -0.012	t=-2.3 (5)	p = 0.07	comb. - artic. = 0.002	t=0.5 (17)	p = 0.61
	artic. - acous. = 0.076	t=4.1 (5)	p < 0.025 *	artic. - acous. = 0.052	t=12.2 (17)	p < 0.025 *
acc′	comb. - artic. = -0.021	t=-2.2 (5)	p = 0.08	comb. - artic. = -0.000	t=-0.1 (17)	p = 0.95
	artic. - acous. = 0.088	t=2.7 (5)	p < 0.04 *	artic. - acous. = 0.045	t=5.2 (17)	p < 0.025 *


A) onset position

B) coda position

Fig. 14. Differences in anticipation metrics between information sources. (A) onset position. (B) coda position. Differences are between combined (comb.), articulatory only (artic.), and acoustic only (acous.) datasets. Horizontal lines show means, boxes show ±1 s.e., and circles show individual participant values for differences in metrics for each speaker/environment. Values close to zero indicate a similar amount of information between sources.

3.4 Articulator-specific information

Analysis of classification accuracy for pared datasets shows that in general no single articulator was a major contributor to category-related information. In other words, there was not a sufficient amount of information in any particular articulator to distinguish all of three categories. The reader should note that this does not entail that some particular articulators are not more informative than others regarding a specific category. The pared datasets were constructed by removing all articulatory dimensions associated with a given sensor, i.e. TT, TB, LL, UL, or JAW. Fig. 14 shows the loss of classification accuracy for pared datasets, averaged over participants. The loss of classification accuracy is obtained for each participant by subtracting the accuracy function for pared data from the accuracy function for the full set of articulatory data. Larger values indicate more information associated with the pared articulator. The top row of panels show accuracy loss calculated over all categories, the bottom three rows show accuracy loss broken down by target category, and the columns correspond to environments.

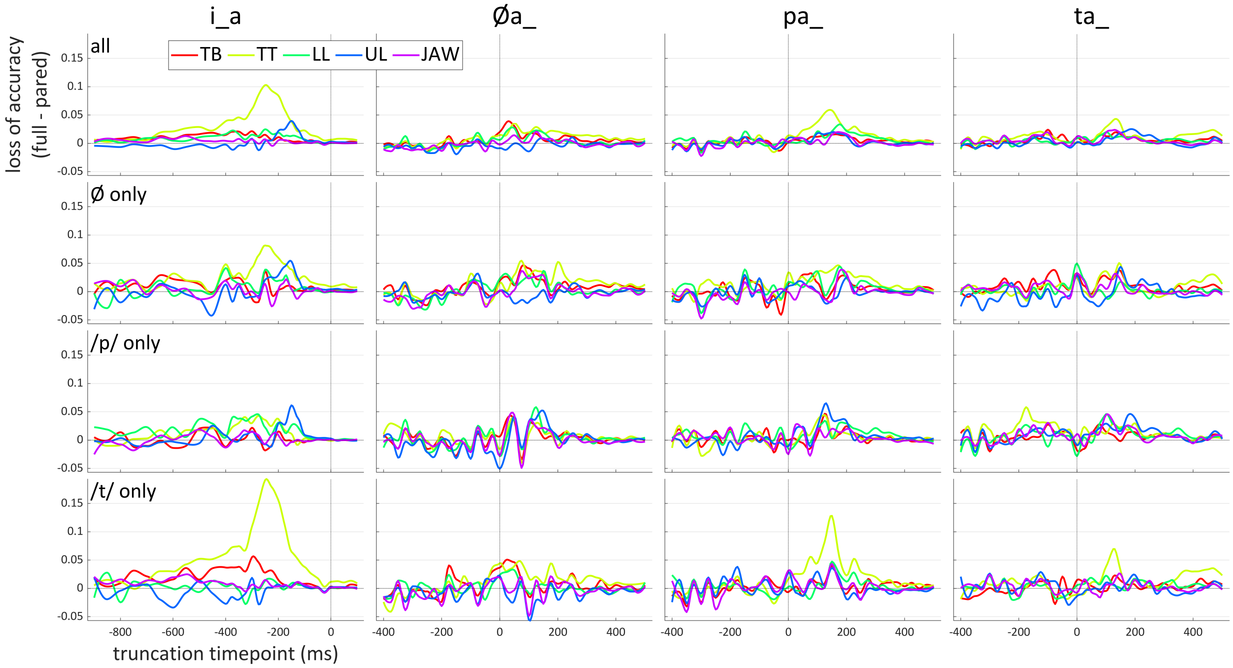


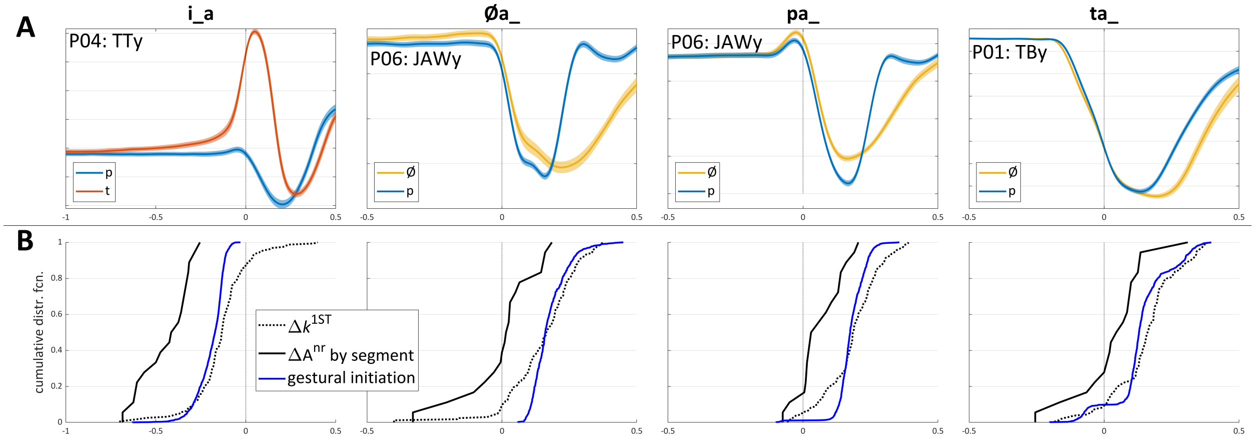
Fig. 15. Loss of classification accuracy for pared datasets, averaged over participants. Larger values indicate more category-related information associated with a given articulator. Top row shows accuracy loss for all categories, bottom three rows show accuracy loss broken down by category. Columns correspond to environments.

The largest accuracy loss when considering all categories was about 10%, for classification of onsets with the TT-pared dataset. As can be seen in the breakdown of accuracy loss by category, this loss was manifested mostly in identification of /t/ in onset position. To a lesser extent /t/ identification in /pa_/ and /ta_/ coda environments was diminished. Notice that no other paring resulted in a substantial loss of accuracy for any given truncation, and that even for truncation times subsequent to gestural initiation, individual articulator paring did not have very large effects on accuracy. In other words, removing signal dimensions related to a single articulator did not generally have a sizeable impact on classification accuracy. This suggests that classification accuracy in the full articulatory dataset makes use of redundant information between articulators.

However, it is evident from the variances of functions in Fig. 15 that the estimates of accuracy loss are somewhat noisy. This is likely because the two functions used in the calculation (pared dataset accuracy and full articulatory dataset accuracy) are both somewhat noisy, each value being an average over 20 repetitions at a given truncation time. Hence more training/test repetitions may be necessary to obtain a clearer picture of articulator-specific contributions to category-related anticipatory information.

3.5 Comparison with standard analysis

It is instructive to compare the signal chopping/network-based estimates of when anticipatory information is present to those that are obtained from a more standard analysis. This comparison demonstrates that the network-based analysis is far more powerful in localizing category-related information. Fig. 16A shows for each environment the most extreme examples of early differences between articulator positions between a pair targets. In each of these examples, the mean vertical or horizontal sensor trajectory is shown along with 95% confidence intervals. The earliest point at which these intervals do not overlap can be viewed as an indirect estimate of when category-related information is present, based on kinematic data. We refer to this metric as Δk^{1ST}. To compare standard analyses with signal chopping/network-based analyses, Δk^{1ST} was calculated for each participant, environment, sensor trajectory, and pair of target categories. Δk^{1ST} is compared with ΔA_{nr}, which is a non-relative version of ΔA estimated by target category. Note that Δk^{1ST} is based on pairwise comparisons of categories while ΔA^{nr} is based on classification of three categories, so the two should not be viewed as directly commensurate. Fig. 16B shows cumulative distribution functions for Δk^{1ST} (dotted line) and ΔA^{nr} (solid black line), along with gestural initiation (blue line). Matrices show the proportion of participants for whom Δk^{1ST} was earlier than the mean gestural initiation time of the relevant categories, broken down by signal dimension and comparison.



	0-p	0-t	p-t	TOTAL
JAWx	0.17	0.17	0	0.11
JAWy	0	0.17	0	0.06
LLx	0.33	0.17	0.17	0.22
LLy	0.17	0.33	0.17	0.22
TBx	0	0	0	0
TBy	0	0.33	0.17	0.17
TTx	0.17	0.17	0.33	0.22
TTy	0	0.50	0.50	0.33
ULx	0	0	0	0
ULy	0.50	0	0.33	0.28
TOTAL	0.13	0.18	0.17	0.16

	0-p	0-t	p-t	TOTAL
JAWx	0.67	0.50	0.17	0.44
JAWy	0.67	0.67	0.17	0.50
LLx	0.50	0.50	0.33	0.44
LLy	0.67	0.17	0.67	0.50
TBx	0.33	0.33	0.33	0.33
TBy	0.67	0.33	0.67	0.56
TTx	0.33	0.50	0.67	0.50
TTy	0.67	0.67	0.33	0.56
ULx	0.33	0.17	0.17	0.22
ULy	0.67	0.17	0.67	0.50
TOTAL	0.55	0.40	0.42	0.46

	0-p	0-t	p-t	TOTAL
JAWx	0.17	0.17	0.17	0.17
JAWy	0.33	0.33	0.17	0.28
LLx	0.17	0.17	0	0.11
LLy	0.33	0.33	0.33	0.33
TBx	0.17	0.17	0.33	0.22
TBy	0.83	0	0.33	0.39
TTx	0.17	0.50	0.67	0.44
TTy	0.50	0.33	0.50	0.44
ULx	0.17	0.17	0.33	0.22
ULy	0.50	0.50	0.67	0.56
TOTAL	0.33	0.27	0.35	0.32

	0-p	0-t	p-t	TOTAL
JAWx	0.17	0.17	0	0.11
JAWy	0.50	0.33	0	0.28
LLx	0.33	0.17	0.17	0.22
LLy	0.33	0.33	0.17	0.28
TBx	0.33	0	0.33	0.22
TBy	0.50	0.33	0.17	0.33
TTx	0.17	0.17	0.50	0.28
TTy	0.17	0.67	0.50	0.44
ULx	0.33	0	0.17	0.17
ULy	0.50	0	0.50	0.33
TOTAL	0.33	0.22	0.25	0.27

Fig. 16. Comparison of signal chopping/network analysis with standard analysis. (A) Examples of early differences in sensor position between targets in a given environment. The participant/articulator channel with the earliest difference for each environment is shown. (B) Comparison of standard analyses with signal chopping/network-based analyses. Plots show cumulative distribution functions for ΔA^{nr} (solid black line), the time of first significant difference between kinematic signals (Δk^{1ST}, dotted line), and gestural initiation (blue line). Matrices show the proportion of participants for whom Δk^{1ST} was earlier than the mean gestural initiation time of the relevant categories.

As is evident from the cumulative distribution functions of Fig. 16B, the network-based measure ΔA^{nr} detects category-related anticipatory information much earlier than the standard approach based on differences in articulator trajectories. It is notable however that even the standard approach is sometimes able to detect differences between categories before gestural initiation. To assess whether these patterns revealed any relations between articulator channels and target categories, contingency analyses of the proportions of participants for whom there were early differences were conducted. Early differences were counted by articulator channel and by category comparison, and were defined as cases in which Δk^{1ST} was less than the mean gestural initiation time for the relevant gestures. Contingency (i.e. χ^2) analyses did not show any significant interactions between channel and comparison. Collapsing over horizontal and vertical dimensions of channels also did not result in any significant interactions. Thus even though there were plenty of statistical differences between sensor channels prior to gestural initiation, no pattern was found regarding which sensor channels were relevant to detecting differences between a given pair of categories.

4. Discussion

The main finding of this study is that there is category-related information in speech signals well before initiation of the articulatory gestures associated with those categories. The analyses presented in Section 3.1 showed that above-chance classification accuracy for a given position (i.e. onset or coda) is possible before the relevant constriction gesture has been initiated. For some participants/environments, onset category information was present up to approximately 350 ms before gestural initiation, and coda information was present up to 250 ms early, during the period of time typically associated with an onset consonant. These findings show that standard AP/TD undergenerates, and provide support for the intentional planning model.

First, lets consider whether there are there any plausible alternative explanations for the main result. One might wonder if the early information is an artefact of the network design or training/testing procedure, since after all, deep networks and LSTMs in particular are very powerful learning mechanisms. This does not seem plausible because when the datasets are truncated early enough, accuracy falls to chance levels. This shows that the classification accuracy cannot be attributed to LSTMs being "very powerful"—the networks still need sufficient information to be successful in classification. Moreover, it is important to remember that the partitioning of data into training and test sets precludes the possibility of spuriously high accuracy from overfitting. The networks never see half of the data (the test set), and accuracies are always calculated on this unseen half. Neural networks can appear to "magically" learn arbitrary mappings between inputs and outputs in training data. This is known as overfitting, and such learning is not generalizable: arbitrary input-output relations learned from training data will not generalize to test data, and often diminish accuracy on test data. In contrast, the accuracies observed on test data represent learning of input-output correspondences that are generalizable. Hence the results should not be dismissed as a consequence of the "magic" of neural networks.

Another possible explanation is that gestural initiation times were mis-estimated and that gestures systematically began earlier than the estimates. Yet the distributions of estimates obtained from the gestural initiation identification procedure (see Section 2.3) were well within expected ranges and were approximately Gaussian for each speaker/position/gesture; moreover, estimates that were inspected visually were always judged correct. Thus it is does not seem plausible to view the results as a consequence of mis-estimated gestural initiations. Although there is some imprecision in the estimates due to the use of an RMS velocity threshold (20% of the maximum RMS associated with the constriction formation), this imprecision is on the order of 10-20 ms (see Appendix: Correspondence between gestural initiation and empirical estimates), and therefore cannot account for the presence of anticipatory information up to several hundred milliseconds before gestural initiation.

Perhaps the data processing procedure is somehow causing information present in later periods of time to spread into earlier periods. For example, could the filtering/smoothing of articulatory data or the windowing for MFCCs somehow be responsible? These explanations do not seem plausible either. The filtering of articulatory data was implemented with a 4th-order Butterworth 10 Hz lowpass on a 400 Hz signal; this filter cannot cause information to spread more than 10 ms (5 samples) backwards in time; moreover, the smoothing window applied to RMS velocity signals was 10 ms. The MFCC windows were larger (30 ms) but the temporal spreading of MFCCs cannot account for the extent to which early information was present, and moreover, the early information was observed in the absence of acoustic dimensions.

Could there be some more indirect explanation? For example, perhaps the orthographic representations of the stimuli were processed differently in a way that influenced the response vowel or the pre-response vowel. Stimuli with /p/ and /t/ codas included the grapheme "o" (*op*, *ot, pop, pot, top, tot*), while stimuli with /Ø/ codas included the graphemic sequence "ah" (*ah*, *pah*, *tah*). Perhaps the percepts of these graphemes induced subtle changes in the articulation of the response vowel or pre-response vowel, thereby creating information in the signal that could lead to early above-chance classification. Or maybe, some of the stimuli were more difficult to process visually than others (perhaps because of variation in familiarity), and this could have had global effects on production due to constraints on resource allocation between visual processing and motor planning. Although such explanations cannot be ruled out, they seem fairly unlikely. They cannot account for any of the more specific effects on anticipatory information which relate to environment (onsetless vs. with-onset) or information source. Moreover, the task is simply not that difficult (error rates were very low), which makes a processing/resource allocation explanation hard to justify.

There were some specific findings related to syllable position and environment that are potentially important. First, it was observed that anticipatory information for an onset category was available earlier than anticipatory information for a coda category (an effect of about 80 ms on average). However, this positional asymmetry should not necessarily be interpreted as evidence that onset gestures are gated in a more leaky manner than coda gestures. The two syllable positions are not directly comparable because the codas were preceded by the vowel /a/ and a varying onset category, whereas the onsets were always preceded by a prolonged /i/. Thus the differences associated with position are confounded with preceding vowel identity, the prolonged nature of the /i/, and the presence/absence of a preceding onset.

Indeed, there were no coda environments in which coda information was detected prior to the average onset initiation time. In the context of the intentional planning model, this suggests that the coda is not sufficiently active to exert an observable influence on the vocal tract prior to initiation of the onset gesture; or, the influence is not detectable with the current design. The design decision to require a prolonged /i/ before the response may also be responsible for the positional asymmetry. The prolonged nature of the /i/ may have made this vowel less speech-like and more like a speech ready posture, which can be adjusted to minimize movement range or other aspects of control (Ramanarayanan et al., 2010). If speech ready postures are modeled as relatively weak gestural forces, they would be less likely than normal gestures to mask the subthreshold effects on intentional fields. Recall from section 1.3 that gestural systems exert forces on intentional planning fields, and that the current target of a tract variable system is the centroid of activation in the corresponding field. Hence when multiple gestures exert forces on a field, the current target depends on the relative strength of those forces. If the prolonged /i/ functions as a speech ready posture, and if speech ready postures are associated with weak forces, this would explain why it is easier in the current experimental design to detect onset category information during /i/ than it is to detect coda category information during preceding /pa/, /ta/, and /Øa/. It is worth mentioning that if no pre-response posture was required, very extensive anticipatory information would most likely be available (Tilsen et al., 2016).

Second, it was observed for targets in coda position that there was more category-related information in the onsetless environment (Øa_) than in the with-onset environments (pa_, ta_), and such information was detectable somewhat earlier (about 50 ms earlier on average). This finding is less equivocal than the position effect (i.e. onset vs. coda), because it does not suffer from the aforementioned confounds. The same explanation offered for the position effect above can be applied to the environment effect here. In the with-onset environments, when an onset consonantal gesture is selected, there are relatively strong forces on intentional planning fields, and these can mask the subthreshold influences of a coda. In the onsetless condition, no onset consonantal gesture is present, and hence there is no masking effect. Under this interpretation, the gating functions for coda gestural forces need not vary by environment; such influences could be constant but hard to detect when onset gestural forces are present. It is worth mentioning, however, that in the "onsetless" environment, speakers did typically exhibit a glottal stop or some adjustment of vocal fold state, as is common in word-initial onsetless environments in English. Thus there is likely a |GLO clo| gesture in this environment. The masking interpretation can still hold, though, because a |GLO clo| gesture does not influence the same tract variables as |LAB clo| or |ALV clo|, and there is no overlap in the articulators recruited.

Another experimental finding is that there were no substantial differences in classification accuracy as a function of target category (Section 3.2). For the relative measures ΔA and acc′, it was the case that for a given position, no subset of categories was more or less accurately classified than any other subset. Only for the non-relativized measure μ_{acc} were such differences observed, and these are expected on the basis of differences in gestural initiation time relative to the reference event to which trials were aligned. The absence of category-related differences in relative measures could be taken to suggest that the mechanism responsible for generating early information is global and not gesture-specific. However, the relevant gestures for /p/ and /t/ (i.e. |LAB clo| and |ALV clo|) are quite similar in that they both (i) have targets which involve an anterior oral closure and pressure build-up, (ii) are coupled to glottal abduction gestures, and (iii) are typologically common and acquired relatively early in development. Perhaps in an experiment with more dissimilar categories—e.g. /t/, /z/, and Ø—gesture-specific differences in anticipation would be observed. Indeed, this sort of question leads to a far more extensive research program in which the composition of the response set and similarity of the relevant categories are manipulated. Another reason to be cautious in interpreting the absence of category-specific accuracy differences is that the training/test procedure is conducted on balanced datasets: the same number of trials of each category is provided in each training set, and so perhaps it is unsurprising that the network learns classification functions that perform approximately the same for each category.

Consistent with the articulatory-acoustic asymmetry hypothesis, it was found that articulatory signals contained more anticipatory category-related information than acoustic signals, while articulatory vs. combined data had similar amounts of information (see Section 3.3). Some degree of caution is warranted in drawing conclusions from these patterns. A fairly reasonable interpretation is that the acoustic information is mostly redundant with the articulatory information, but not vice versa. Articulatory data might be expected to contain more information because of the synergistic nature of the control of vocal tract geometry. For example, the jaw might be positioned differently in anticipation of different targets, while lingual and labial positions are compensatorily adjusted such that there is no effective acoustic difference. The anticipatory effect would then be observable only in articulatory data.

The redundancy of acoustic signals may be a sensible interpretation of the findings because patterns of acoustic energy generated in speech are necessarily determined by physical mechanisms from the state of the vocal tract (neglecting variation in atmospheric properties and recording equipment/environment). However, there are numerous speech-related systems—namely pulmonic, laryngeal, and velaric—that are not represented in the articulographic data collected in this experiment; these systems always interact with oral vocal tract geometry to generate patterns of acoustic energy. It follows that the acoustic data might in fact be expected to contain information that is not present in articulatory signals, and hence might not be a redundant source of information. Note however, that this would only be the case if the articulatory gestures in the classification set are associated with differences in the pulmonic/laryngeal/velaric influences on the acoustic signal. For /p/ and /t/ it is unclear to what extent such differences might exist.

A more important caveat in interpreting the effects of information source is that the particular dimensionality reduction used to construct acoustic signals may not be optimal, or may obscure more information than the dimensionality reduction imposed on articulatory signals. Consider that there are many experimenter degrees of freedom (i.e. hyperparameters) involved in constructing the signals: the MFCC coefficient matrices are determined by eight parameters (see Section 2.2), most of which are continuous—this implies a very large space of analyses which cannot be practically explored. It is therefore nearly impossible to know if the particular MFCC parameters used here are the optimal ones. For example, the choices to use 30 ms windows and 5 ms frame steps were made based on typical MFCC analyses, but these are unlikely to be optimal for the current purposes. Indeed, it is sensible to ask whether MFCCs are the best choice—why not spectrograms or even the raw acoustic signal? This sort of question can be addressed in future research by comparing classification accuracies obtained from various forms of network input, but there is an important consideration to take into account. The reader should note that the use of time-frequency representations (i.e. cepstral coefficients or spectrograms) can be partly motivated by the fact that it is straightforward to control the reduction of temporal and spatial dimensionality via conversion to the spectral/cepstral domain. In contrast, the raw acoustic signal is much higher in dimensionality and so training LSTMs directly on acoustic signals would require an enormous increase in processing power. One can therefore anticipate that in the future, with more powerful systems, this will be possible. Nonetheless, for the reasons above, the observed differences do not unequivocally support the hypothesis of articulatory-acoustic information asymmetry.

The main finding from analyses of articulator-paring was that no single articulator contributes a substantial amount of category-related information. This result may not be surprising, for two reasons. First, recall that category-related information is information that allows the networks to accurate classify all three categories; hence we would not necessarily expect that any particular articulator would be useful for discriminating between /p/, /t/, and Ø categories, even though a particular category may be more or less strongly associated with a particular articulator. Second, as explained in section 1.2, the AP/TD model predicts that any given oral constriction gesture will influence tract variables and articulators which are not directly associated with the gesture, given biomechanical coupling of articulators with the jaw. Hence it is expected that the information necessary to classify speech categories is distributed across a variety of articulatory dimensions. Importantly, the result does entail that there is no category-specific information in individual articulator channels; more likely, it suggests that the information in any particular channel is redundantly present in other channels. The articulator with the largest accuracy loss when pared was the tongue tip (TT) in onset position, which reduced overall accuracy by about 10% for truncations in the vicinity of gestural initiation. The observation that classification accuracy does not depend strongly on any single articulator suggests that there is redundant information between articulators. Of course, the next logical steps in assessing articulator-specific information are to pare combinations of two, three, and four articulators (these have 10, 10, and 5 unique combinations respectively). This could be taken further by paring horizontal vs. vertical signal dimensions, as well as derivatives. In the future when more computing power is available, such analyses might be commonplace.

5. Conclusion

This study shows that the AP conception of how and when a gesture influences the state of the vocal tract must be revised. Specifically, one cannot view this influence as an event which is discretely bounded in time; nor should one assume that the beginning of the influence of a gesture is highly correlated with the initiation of a substantial movement. Some gestural planning processes must intervene between gestural retrieval and movement initiation, and under the right circumstances these processes are detectable in behavioral data. Accordingly, movement initiation should not be treated as an index of when those planning processes begin, and perhaps not even as an indirect correlate of that beginning. As predicted by the intentional planning mechanism in the Selection-coordination-intention framework (Tilsen, 2018, 2019), gestures which have not yet been initiated can have an influence on the state of the vocal tract when the gestural force gating function is relatively leaky. Tilsen (2019) proposed that this leakiness could be a precursor to the emergence of consonant harmonies. If so, on the basis of typological asymmetries in such harmonies, the model predicts that different gestures should exhibit different degrees of anticipatory effect on the vocal tract, and that these differences might correlate with the organization of the gestural inventory of a language or with the statistical distributions of gestures.

Another possibility worth investigating is whether the leakiness of the gestural force gating function can be controlled with experimental manipulations. This could be accomplished by drawing attention to one of the specific categories in a response set, for example, by depicting just the grapheme “p” in a bold red font and requiring the participant to keep track of how many times they see “p”; heightened attention to a particular category might cause the corresponding gesture to be more highly excited when it is in the subthreshold state, resulting in a stronger influence on the vocal tract. Some related ideas are to require a speeded response only when the target word contains some particular category, or to instruct speakers to hyperarticulate just one particular category. All of these manipulations would be predicted to increase the degree of anticipation.

Perhaps the time-course of subthreshold intentional planning can be controlled as well. For example, in the current experiment, there was a two second period in which the target response was displayed prior to the go-cue, and participants produced the pre-response vowel /i/ during this period. Most of the participants only began to exhibit above-chance classification of the onset several hundred seconds before gestural initiation, and so it does not appear to be the case that subthreshold intentional planning had an influence on the vocal tract throughout the entirety of the pre-response vowel (although the data processing and network training procedures are not optimal and hence information is likely to be present somewhat earlier). It stands to reason that in an unprepared, speeded response task, where the target-stimulus and go-cue are simultaneous and where the response occurs shortly after retrieval, there should be almost no anticipatory information before movement initiation. The intentional planning model thereby makes strong predictions regarding how the relative timing of the target stimulus and go-cue should affect the time course of anticipatory information.

Another topic for future investigation is whether category-related effects can be dissociated from word-related effects. The current study partly conflates these two sources of effect and did not control the wordhood of the target forms. In future experiments, a target set could be designed to address interesting questions which relate to lexical knowledge. For example, consider the homophone set *time*, *thyme*, and *tyme* (a novel word); signal chopping might be useful for characterizing the temporal and spatial distribution of acoustic/articulatory information which differentiates these forms. However, it is worth considering that lexical item-specific effects may be diminished in experimental paradigms which elicit the same items repeatedly, because the items become highly familiar on the timescale of the experimental session. Perhaps corpus data from conversational speech can be used instead, but it is not yet known whether the method can be successful when applied across multiple speakers and environments.

Finally, a very general contribution of this paper is the development of signal chopping, a method for localizing information in space and time. Signal chopping can be used with almost any sort of analysis procedure, although machine learning or deep learning are particularly useful when the data are high-dimensional. The techniques employed here can be applied to a wide variety of contexts and may reveal previously unrecognized patterns. It is noteworthy that although we focused on anticipatory information in this study, a similar analysis could be conducted in order to determine how long category-related information perseverates in articulatory and acoustic signals. Even more generally, the signal chopping procedure could be used to construct a time vs. time-scale analysis of category-related information, where the centers and sizes of analysis windows are varied systematically (similar to wavelet analyses). The resulting accuracies would represent category-related information density as a function of time and time-scale. An important benefit of approaching our analyses in this way is that it allows us to make fewer assumptions regarding which aspects of speech signals are the relevant ones.

Acknowledgements

I would like to thank Abby Cohn, Martin van Schjindel, and members of the Cornell Phonetics Lab for discussions of this research, and Bruce McKee for equipment-related assistance.

Appendix: Correspondence between gestural initiation and empirical estimates

To establish an expected degree of correspondence between gestural initiation and empirical estimates thereof, parameters of a Task Dynamic model were optimized to fit the LA timeseries for the onset consonantal constriction for each onset [p] trial of the experiment. The optimized activation functions and empirical tract variables were then analyzed to determine the expected deviation between empirical estimates of gestural initiation and the optimized theoretical values. With step function gestural activation, the mean difference between empirical estimates and optimized gestural initiations was 7.6 ms with a standard deviation of ±2.5 ms. With ramped activation functions, which provide better fits to tract variable time series, the mean difference between empirical estimates and optimized gestural initiations was -12.9 ms with a standard deviation of ±3.1 ms. Details of the model, optimization, and analysis are described below.

A sometimes underappreciated virtue of the AP/TD framework is that there is an explicit mathematical model which relates theoretical entities, gestures, to empirical observations, tract variables. The relation is specified implicitly by equation (1) below:

(1) $\ddot{x} = \beta(t)\dot{x} + k(t)\big(x - T(t)\big)$

The variable x is a tract variable, such as lip aperture (LA), and $\beta(t)$, $k(t)$, and $T(t)$ are time-varying damping, stiffness, and target (i.e. equilibrium). The standard implementation of the Task Dynamic model (Saltzman & Munhall, 1989) is designed to allow for multiple gestures to simultaneously influence the target of the same tract variable, and to capture effector-level interactions between gestures which specify targets for different tract variables. For current purposes these additional complications are unnecessary and hence we can simplify the description of the time-varying quantities. The target in the simplified model is specified as in (2):

(2) $T(t) = a(t)\hat{T} + \big(1 - a(t)\big)x_0$

The variable $a(t)$ is a time-varying gestural activation normalized to a range of [0,1]. Equation (2) holds that the current target is a weighted combination of the gestural target and a neutral attractor target, with the weights being the gestural activation and activation of a competitively blended neutral attractor. In the optimizations conducted here, the target of neutral attractor is assumed to be the initial value of the tract variable x_0, defined as the extremum of LA which most immediately precedes the movement. Following (Saltzman & Munhall, 1989) we will assume critical damping, so $\beta(t) = 2\sqrt{k(t)}$, and we will furthermore assume that the neutral attractor and gesture have constant stiffness, $\hat{k}$. Hence Equation (1) can be expressed as in (3):

(3) $\ddot{\mathrm{x}} = 2\sqrt{\hat{k}}\dot{\mathrm{x}} + \hat{k}\big(\mathrm{x} - \big[a(t)\hat{T} + \big(1 - a(t)\big)x_0\big]\big)$

When gestural activation $a(t)$ is modeled as a step function, the imposition of linear stiffness results in somewhat poor fits of empirical data. As discussed in Sorensen & Gafos (2016), one remedy for this is ramped activation, where the gestural activation is allowed to increase gradually from 0 to 1. We parameterize the duration of this increase as δ_{ramp} and define $a(t)$ as in Equation (4) below. Note that since we are fitting only the constriction phase of movements, we do not need to model the deactivation of the gesture.

$$(4)\quad \mathrm{a}(\mathrm{t}) = \begin{cases} \min\left(\frac{\mathrm{t}-\mathrm{t}_{\mathrm{init}}}{\delta_{\mathrm{ramp}}}, 1\right), & \mathrm{t} \geq \mathrm{t}_{\mathrm{init}} \\ 0, & t < t_{init} \end{cases}$$

The parameter t_{init} is the gestural intitiation time. Equation (4) states that gestural activation is zero before t_{init}, and the *min* function enforces a ceiling of 1 on gestural activation.

For each onset [p] trial of the experiment we optimize the parameters t_{init}, $\hat{k}$, $\hat{T}$, and δ_{ramp} to minimize the mean squared error between the model output and a portion of the empirical LA timeseries. The selected portion was from the time of the minimum preceding the velocity extremum, to the time of the maximum following the velocity extremum. To enforce step activation, δ_{ramp} is fixed at 0. The empirical estimate of gestural initiation is obtained using a 20% velocity criterion: it is the last time at which the LA timeseries is below 20% of the speed maximum associated with the |LAB clo| gesture, prior to the time of the speed maximum.

In most cases, the resulting optimizations had very low mean square errors (MSE): 75% of the MSEs were below 0.01 mm with activation ramping, i.e. less than 1/100 of a millimeter of error per sample. However, in some cases the empirical trajectories cannot be well-fit by the model and have high MSE. These cases are ones in which the preceding minimum that was used for selecting the empirical trajectory occurs too early (indeed, this is why a 20% velocity threshold is used in empirical estimates). In such cases, the model is trying to minimize MSE over a period of time which extends well before substantial evidence of movement (i.e. before a local change in LA which as a large velocity extremum). In order to ensure that gestural initiation estimates faithfully reflect constriction movements, only the optimizations associated in the lower three quartiles of MSE were analyzed.

The discrepancies observed between gestural initiation and empirical estimates in this dataset were quite small. With step function gestural activation, the mean difference between empirical estimates and optimized gestural initiations was 7.6 ms with a standard deviation of ±2.5 ms. With ramped activation functions, which provide better fits of tract variables, the mean difference between empirical estimates and optimized gestural initiations was -12.9 ms with a standard deviation of ±3.1 ms. The means and standard deviations of the discrepancies and the optimized parameters for each participant are shown in Table A.1.

Table A.1. Discrepancies between optimized gestural initiation and empirical estimates

	discrepancy (ms) (mean, st. dev.)	**target – x_0 (mm) (mean, st. dev)**	**stiffness × Δt (mean, st. dev)**	**initiation time (s) (mean, st. dev)**	**ramp duration (s) (mean, st. dev)**
step activation:					
P1	8.0 (2.4)	-7.7 (1.0)	1.9 (0.4)	0.0287 (0.0076)	
P2	9.1 (2.3)	-10.3 (1.7)	1.6 (0.3)	0.0313 (0.0067)	
P3	7.4 (1.3)	-12.1 (1.8)	1.9 (0.6)	0.0246 (0.0057)	
P4	7.3 (3.2)	-9.4 (3.2)	1.9 (1.6)	0.0291 (0.0106)	
P5	6.9 (3.8)	-13.0 (2.1)	1.2 (0.7)	0.0286 (0.0091)	
P6	7.0 (2.7)	-8.6 (1.3)	1.9 (0.4)	0.0250 (0.0054)	
ramped activation:					
P1	-11.1 (1.6)	-7.7 (1.0)	5.1 (1.0)	0.0095 (0.0081)	0.0685 (0.0077)
P2	-12.3 (1.7)	-10.3 (1.7)	4.1 (0.9)	0.0099 (0.0072)	0.0750 (0.0076)
P3	-12.5 (1.9)	-12.1 (1.8)	4.3 (1.2)	0.0048 (0.0053)	0.0652 (0.0080)
P4	-13.5 (6.4)	-9.3 (3.3)	4.2 (1.7)	0.0078 (0.0113)	0.0676 (0.0157)
P5	-16.5 (3.1)	-12.8 (2.0)	2.4 (1.1)	0.0062 (0.0108)	0.0726 (0.0116)
P6	-12.4 (1.8)	-8.6 (1.2)	4.3 (1.2)	0.0055 (0.0065)	0.0641 (0.0060)

The above analysis of correspondence between gestural initiation and its empirical estimate can be generalized to other tract variables and more realistic gestural contexts. Here we focused on |LAB clo| gestures in onset position. The analysis can be extended to |ALV clo| gestures, but this is slightly more complicated because it is preferable to transform the horizontal and vertical components of the TT sensor to a principal component dimension. The analysis can also be extended to coda position without loss of generality. However, there is one complicating factor that is worth mention. In the above simulations we assumed just one active gesture influences the relevant tract variable, LA. More realistically, other gestures, such as the preceding and following vocalic gestures—|PAL [i]| and |PHAR [a]|—will overlap with |LAB clo| and will thus perturb LA indirectly via their influences on the JAW articulator. Assuming that these influences are either fairly constant or are relatively small, the estimated discrepancies between gestural initiation and empirical measures will be reliable.

Appendix: Network details

The parameter spaces that define network structure and training procedures are very large, and hence it is not practical to obtain systematic motivations for most design choices. For the most part, the design choices in the current study were made on the basis of (i) exploratory tests, (ii) examples in function documentation, and (iii) logistical considerations. The logistical considerations include the computer memory available for training and trade-offs between network size and training time. All network training was conducted on a single GPU (NVIDIA GTX 1080, 8 GB RAM, 2560 Cuda cores), using the Matlab Deep Learning toolbox. Each network took on average roughly 40 s to train. Since 136,320 networks were trained, this amounted more than 1500 hours (about 63 days) of GPU time. Prior to this, exploratory tests of various network structures and training procedures were conducted with the aim of identifying parameters that successfully classified segments in non-truncated data. The exploratory tests identified design parameters that reached nearly 100% accuracy in non-truncated sequences. However, these exploratory tests were not systematic and were not conducted on the entire dataset. Thus it almost certain that the network structure and training procedures were not optimized for classification of onset and coda segments. Thus, the accuracies obtained are necessarily *underestimates* relative to chance accuracy.

The classification network contained three layers of 400 bidirectional LSTM units. Bidirectional LSTM units were used rather than unidirectional LSTM units because the former are expected to provide better accuracy in classification. Bidirectional LSTM units have access to information both from past and future input sequence states, whereas unidirectional LSTM units have access only to information from current and past states. The choice to use three layers rather than one or two was based on superior performance in exploratory testing. The choice to use 400 units per layer was based on memory limitations that arose in training. Each LSTM layer was followed by a dropout layer with 40% dropout proportion. The dropout layer diminishes overfitting in the training phase. The third biLSTM-dropout layer was followed by a fully connected layer, a softmax layer, and a classification layer with a cross-entropy loss function.

The network training procedure used an adaptive moment estimation optimizer (Adam) with the following parameters. Gradient threshold: 1.0; initial learning rate: 0.0005; L2 regularization 0.0005. These parameters were chosen mostly on the basis of examples in documentation. The optimization was limited to a maximum of 200 epochs, and a mini-batch size of 24 was used. Note that an iteration is one step of the optimization algorithm (including weight updates) applied to a mini-batch, and an epoch is a full pass of the training algorithm over the training set. The order of sequences (and hence their assignment to mini-batches) is shuffled every epoch. All sequences are the same length in a given training run, so no padding is necessary. The validation frequency and patience were 5 and 20 iterations, respectively. This entails that the validation accuracy is calculated after every 5 iterations. If the loss function on the validation set does not decrease over 20 iterations then the network training stops. Note that although the validation data can determine when the training stops, it does not play any role in determining how network weights are updated in the optimization process.

References

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*(4), 591–627.

Browman, C., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. *Between the Grammar and Physics of Speech*, 341–376.

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

Browman, C., & Goldstein, L. (1995). Gestural syllable position effects in American English. *Producing Speech: Contemporary Issues*, 19–33.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, *24*(2), 209–244.

Clements, G. N., & Keyser, S. J. (1983). CV phonology. A generative theory of the syllabe. *Linguistic Inquiry Monographs Cambridge, Mass.*, *9*, 1–191.

Cohen-Goldberg, A. M. (2012). Phonological competition within the word: Evidence from the phoneme similarity effect in spoken production. *Journal of Memory and Language*, *67*(1), 184–198.

Cohn, A. C. (1990). *WPP, No. 76: Phonetic and Phonological Rules of Nasalization*.

Davis, S. (1989). Cross-vowel phonotactic constraints. *Computational Linguistics*, *15*(2), 109–110.

Fudge, E. (1987). Branching structure within the syllable. *Journal of Linguistics*, *23*(2), 359–377.

Gafos, A. I. (1999). *The articulatory basis of locality in phonology*. Taylor & Francis.

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word*, *6*(2), 162–181.

Grosvald, M. (2009). Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation. *Journal of Phonetics*, *37*(2), 173–188.

Hansson, G. Ó. (2001). Consonant harmony: Long-distance interaction in phonology. *UC Publications in Linguistics*.

Hawkins, S., & Nguyen, N. (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset/l/in English. *Journal of Phonetics*, *32*(2), 199–231.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Kawamoto, A. H., Liu, Q., Mura, K., & Sanchez, A. (2008). Articulatory preparation in the delayed naming task. *Journal of Memory and Language*, *58*(2), 347–365.

Krause, P. A., & Kawamoto, A. H. (2019a). Anticipatory mechanisms influence articulation in the form preparation task. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(3), 319.

Krause, P. A., & Kawamoto, A. H. (2019b). Nuclear vowel priming and anticipatory oral postures: Evidence for parallel phonological planning? *Language, Cognition and Neuroscience*, 1–18. https://doi.org/10.1080/23273798.2019.1636104

Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *25*(2), 187–205.

Marin, S., & Pouplier, M. (2010). Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control*, *14*(3), 380–407.

McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, *17*(2), 207–263.

Mücke, D., Tilsen, S., & Hermes, A. (2020). The elephant-in-the-room: Incongruencies of phonological theory and phonetic measurement. *Phonology*.

Nam, H. (2007). Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure. In *In J. Cole & J. I. Hualde (Eds.), Laboratory phonology* (Vol. 9, pp. 483–506). Walter de Gruyter.

Ohala, J. J. (1993). The phonetics of sound change. *Historical Linguistics: Problems and Perspectives*, 237–278.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, *41*(2), 310–320.

Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. *Proceedings of NELS*, *23*, 367–381.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., & Schwarz, P. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Ramanarayanan, V., Byrd, D., Goldstein, L., & Narayanan, S. S. (2010). Investigating articulatory setting-pauses, ready position, and rest-using real-time MRI. *Eleventh Annual Conference of the International Speech Communication Association*.

Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance; Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 307.

Recasens, D. (1987). An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences. *Journal of Phonetics*, *15*(4), 299–312.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382.

Sethna, J. (2006). *Statistical mechanics: Entropy, order parameters, and complexity* (Vol. 14). Oxford University Press. https://books.google.com/books?hl=en&lr=&id=O09uBAAAQBAJ&oi=fnd&pg=PP1&dq=fluctuation+dissipation+theorem+sethna&ots=BIDLjA84Wq&sig=Sy-8wmsGjxip8sZRY_vLSkGnsAM

Shannon, C. E. (1948). *A Mathematical Theory of Communication*.

Sorensen, T., & Gafos, A. (2016). The gesture as an autonomous nonlinear dynamical system. *Ecological Psychology*, *28*(4), 188–215.

Tilsen, S. (2007). Vowel-to-vowel coarticulation and dissimilation in phonemic-response priming. *UC Berkeley Phonology Lab 2007 Annual Report*, 416–458.

Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics*, *37*(3), 276–296.

Tilsen, S. (2013). A Dynamical Model of Hierarchical Selection and Coordination in Speech Planning. *PloS One*, *8*(4), e62800.

Tilsen, S. (2014). Selection-coordination theory. *Cornell Working Papers in Phonetics and Phonology, 2014*, 24–72.

Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics*, *55*, 53–77.

Tilsen, S. (2017). Exertive modulation of speech and articulatory phasing. *Journal of Phonetics*, *64*, 34–50.

Tilsen, S. (2018). *Three mechanisms for modeling articulation: Selection, coordination, and intention* (Cornell Working Papers in Phonetics and Phonology 2018).

Tilsen, S. (2019). Motoric mechanisms for the emergence of non-local phonological patterns. *Frontiers in Psychology*, *10*, 2143.

Tilsen, S., Spincemaille, P., Xu, B., Doerschuk, P., Luh, W.-M., Feldman, E., & Wang, Y. (2016). Anticipatory Posturing of the Vocal Tract Reveals Dissociation of Speech Movement Plans from Linguistic Units. *PLoS ONE*, *11*(1), e0146813. https://doi.org/10.1371/journal.pone.0146813

Walker, R. (2011). Nasal harmony. *The Blackwell Companion to Phonology*, 1–28.

Whalen, D. H. (1990). Coarticulation is largely planned 7/3. *Journal of Phonetics*, *18*, 3–35.