Statistical identification of articulation constraints in the production of speech [⋆]

Philip J.B. Jackson [*] and Veena D. Singampalli

Centre for Vision, Speech and Signal Processing, University of Surrey, UK.

Abstract

We present a statistical technique for identifying critical, dependent and redundant roles played by the articulators during production of English phonemes using articulatory (EMA) data. It identifies a list of critical articulators for each phone based on changes in the distribution of articulator positions. The effect of critical articulation on dependent articulators is derived from inter-articulator correlation. Articulators unaffected or not correlated with the critical articulators are regarded as redundant. The technique was implemented on 1D and 2D distributions of midsagittal articulator coordinates, and the results of this data-driven approach are analyzed in comparison with the phonetic descriptions from the IPA chart. The results using the proposed method gave a closer fit to measured data than those estimated from IPA information alone and highlighted significant factors in the phoneme-to-phone transformation. The proposed algorithm was evaluated against an exhaustive search of critical articulators, and found to be as effective as the exhaustive search in modeling phone distributions with the added advantage of faster execution times. The efficiency of the approach in generating a parsimonious yet accurate representation of the observed articulatory constraints is described, and its potential for applications in speech science and technology discussed.

Key words: critical articulator, speech production model, articulatory gesture, coarticulation

[⋆] This research formed part of the DANSA project funded in the UK by EPSRC (GR/S85511/01), http://personal.ee.surrey.ac.uk/Personal/P.Jackson/Dansa/.

[*] Correspondence: Dr Philip Jackson, CVSSP, University of Surrey, Guildford, Surrey GU2 7XH, UK. Tel.:+44 1483 686044; Fax: +44 1483 686031.

Email address: `p.jackson@surrey.ac.uk` (Philip J.B. Jackson).

1 Introduction

An accurate model of speech articulation is important for understanding its production and perception (Meister et al., 2007; Wilson et al., 2004) and for integration into speech technologies (King et al., 2007). Yet coarticulation remains one of the main problems faced in speech research. In the production of speech, e.g., from a specified sequence of phonemes, coarticulation spreads their influence across the utterance so that substitution of one phoneme for another (or the change of one distinctive feature) alters it not only within the corresponding phone segment but throughout the neighbouring segments. The human speech articulators (jaw, tongue, lips, etc.) have limited freedom to move, interconnections, stiffness, damping and inertia. Speech gestures are thus planned in a coordinated sequence, controlled by intrinsic and extrinsic muscles, and are relatively slow and overlapping. As a result, it is difficult to match linguistic units with acoustic data and to account for all kinds of observed variability in articulation. There have been many acoustic, pseudo-articulatory and articulatory approaches to modeling the spatio-temporal effects of coarticulation. From observations of acoustic data, Lindblom (1963) studied the formant target undershoots of vowels in CVC occurrences as a function of vowel duration and consonantal context across different speaking styles. His locus equation approach has been used to quantify the context sensitive coarticulatory effects by many since. Öhman (1966) investigated coarticulatory effects in VCV utterances, which he explained in terms of a consonantal gesture superimposed on a continuous vocalic one. Those articulators not actively involved in producing the consonantal gesture (i.e., not critical) were most influenced by vowel context. Öhman's (1967) model used static articulatory configurations as idealized target vocal tract shapes and dynamic functions to explain the degree of excursion of articulatory movements along with a function to model coarticulation. Although this approach recognizes the different reaction speeds of the intrinsic and extrinsic muscle groups that drive these gestures, the many assumptions and parameters employed in his model had to be set manually, which does not allow for cross-validation or objective optimization.

Binary and discrete features

Coarticulation theories have tended to focus on one or other of the two main stages in the linguistic-to-acoustic realization. The feature spreading theory considers effects in the planning stage, converting phoneme sequences into distinctive articulatory features, and co-production theory deals with the motor control and physical dynamics of the articulators. In the feature spreading approach, a distinctive set of bipolar phonetic features encode the place, manner and voicing information of a phone (Chomsky and Halle, 1968; Fant, 1969), while non-critical features are left unspecified. Anticipatory coarticulation has been modeled as a spread of critical features to unspecified segments, from right to left (Daniloff and Hammarberg, 1973; Moll and Daniloff, 1971; Henke, 1965). Feature spreading is blocked at the next specified segment. Based on the features, the articulators were given spatio-temporal targets which were assumed to be invariant although possibly not reached due to physical smoothing. One alternative to fixed targets is the window model of Keating (1988), in which a range of articulatory values are allowed for each segmental feature. Some speech recognition systems have been inspired by the binary phonetic feature concept as a means for incorporating articulatory information (Kirchhoff, 1999; Metze and Waibel, 2002; Frankel et al., 2004; Eide, 2001; Koreman et al., 1998). However, the binary features do not describe how tight each constraint is. In reality, the unspecified segments can be partially affected or unaffected by the spreading feature, so the extent of anticipatory coarticulation is not well explained by this theory.

Bladon and Al-Bamerni (1976) moved from a binary specification to discrete coarticulation resistance (CR) for explaining cross-consonantal coarticulation effects in VCV contexts. The CR quantifies the consonant's resistance to V-V coarticulation. Though simple, modeling coarticulation using a binary or discrete set of non-overlapping features for each phone segment has many drawbacks. The time units are discrete segments, which imply synchronous feature boundaries. Articulation is a continuous process, in time and space, and discrete features fail to represent even ideal articulatory configurations adequately. For modeling real speech production, they are poorly suited.

Synchronous vs. asynchronous models

Some researchers have used discrete articulatory features, where the vocal-tract configuration for a given phone was represented as a set of quantized articulator positions which were then mapped to the states of a hidden Markov model (Deng and Sun, 1994; Erler and Freeman, 1996; Richardson et al., 2000). The overlap of the discrete gestures was modeled by spreading the quantization values to the articulatory dimensions not crucially involved in the production of a speech sound. Here, though the time domain was in discrete frames, the feature boundaries were asynchronous. The values for each phone were set manually from phonetic knowledge; in our approach, values are determined directly from measured data.

Articulator movements have been recorded in many ways, e.g., using EMA (electro-magnetic articulograph), X-ray (Westbury et al., 1994; Soquet et al., 1999; Wrench, 2001) and tagged MRI (Parthasarathy et al., 2007). Efforts to capture in equations the dynamic movements of articulators towards phone specific goals have led to gestural approaches (Browman and Goldstein, 1986; Saltzman and Munhall, 1989; MacNeilage, 1970; Liberman, 1970). Articulatory gestures are associated with an intrinsic temporal structure that allows for continuous and asynchronous movements. Overlap from co-production of gestures results in coarticulation. Research has developed equations to model muscle behaviour and articulator kinematics (Coker, 1976; Ostry et al., 1996; Dang and Honda, 2004). In the dominance function approach to modeling coarticulation (Löfqvist, 1990; Cohen and Massaro, 1993), each segment is viewed as a bundle of gestures, one for each articulator, linked over time by exponential functions that vary in duration and magnitude. In these approaches, the dynamic functions for the activated gestures are prescribed according to phonetic rules. Another use of rules is in determining gesture priorities. A scale proposed for the tongue is degree of articulatory constraint (DAC) (Recasens et al., 1997; Recasens and Pallarés, 1999), which describes the extent to which a consonant or vowel constrains tongue dorsum motion in VCV contexts: the higher the value, the larger the resistance to coarticulation. Mermelstein used three levels to rank how critical an articulatory gesture was to a given phone (Mermelstein, 1973). The idea of configurations critical in production of a sound (crucial points) has been used in the study of coarticulation (Dang et al., 2004). The crucial articulator was defined as being resistant to contextual effects and having maximum coarticulatory influence on its neighbours. The critical or non-critical roles of the articulators was specified from phonetic rules. The present work aims to determine articulatory roles from quantitative measurements.

Statistical techniques have long since taken over from rule-based approaches for dealing with coarticulation in automatic speech recognition (ASR). Dang et al. (2005) proposed a descriptive statistical model based on Öhman's view of coarticulation, and incorporating features like CR and DAC that were estimated from the articulatory data in their model. Using an HMM to provide a probabilistic representation of articulatory target distributions was proposed by Bakis (1991), which is an extension of Keating's rectangular windows (Keating, 1988). Other techniques for generating smooth trajectories from probabilistic descriptions include dynamical models (Richards and Bridle, 1999) and trajectory HMMs (Tokuda et al., 2007). Context-sensitive effects on those distributions from articulator acceleration have also been modeled (Blackburn and Young, 2000). Statistical models are powerful in making good use of available articulatory data to describe phone-sequence characteristics, but fail to identify the cause of the constraints offered by the speech production system and hence are not parsimonious. State-of-the-art TTS and ASR systems tend to use ever longer units and models to accommodate coarticulatory effects without explicit knowledge of the articulatory constraints that convert a phoneme string into speech.

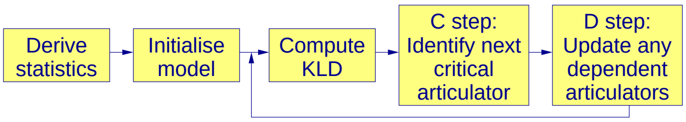

Derive statistics
Initialise model
Compute KLD
C step: Identify next critical articulator
D step: Update any dependent articulators

Fig. 1. Critical articulator identification process.

We present a statistical approach for identifying constraints in the articulatory domain considering the roles that articulators play during speech production. We categorize the roles as critical, dependent and redundant. If an articulatory gesture or movement plays a crucial role in the production of a phone, it is considered to be *critical*. As in previous work, the critical articulators were associated with smaller variance when compared with the non-critical articulators (Papcun et al., 1992; Frankel and King, 2001). Here, we considered that an articulator can also be critical if characterized by a significant shift in mean position with respect to its neutral average state. Furthermore, the form of articulatory variation in space could be specific to production of a certain phone, such as when the tongue impinges on the roof of the mouth changing the correlation with respect to its average, so we considered changes in covariance too. We defined a *dependent* articulator as one whose position follows from the influence of a critical articulator because of the bio-mechanical correlations between them. In gestural theory, this kind of articulatory gesture is termed as a passive gesture following its correlation with the active tract variable (Saltzman and Munhall, 1989), and is explained using the concept of controlled and uncontrolled manifolds. A *redundant* articulator is free to move and, as its position does not affect the phone's production in a critical way, coarticulation effects tend to be strong inducing a full range of movements. The algorithm we propose makes use of correlations (amongst articulators and correlated movements of each articulator in space) to identify the critical, dependent and redundant articulators for every phone. The model is entirely data driven and generates parsimonious representations of the articulatory target configurations for each phone, based on phone transcriptions of the synchronised audio signal. Our original approach to the problem of coarticulation offers an experimental and quantitative validation of most of the traditional IPA phoneme descriptions. The next sections present the method, phonological analysis of results, evaluation and discussion of the possible applications of the algorithm, and conclusion.

2 Method for identification of articulatory roles

We present a statistical algorithm for identifying the critical, dependent and redundant roles of articulators during speech production of English phonemes. Ideally, the algorithm should accurately identify what constraints each articulator experiences throughout the production of a given speech utterance. These constraints should be related to phoneme units, specified quantitatively and derived from realistic speech data. Also, for practical reasons, it would be advantageous to find an efficient algorithm. The work here is based on the statistical analysis of quantitative articulatory measurements obtained by electro-magnetic articulography (EMA). As shown in Figure 1, our iterative algorithm has five main steps, described in detail later in this section: estimation of articulatory statistics, model initialisation, distance calculation, identification of next critical articulator (C step), and update of dependent articulators (D step). Section 3 examines the algorithm's accuracy, Section 4 carries out a phonological analysis and Section 5 evaluates the proposed method against a 'brute force' approach.

We take samples of articulator coordinates (horizontal and vertical) at the middle of each phone label to provide an approximation of the target distribution for that phone. The complete set of these samples defines the grand distribution for each articulator coordinate. The corresponding means and variances are used to form Gaussian pdfs (probability density functions) to represent the phone and grand distributions. Similarly, we compute the inter-articulator correlations from these data to give grand correlations. For 1D and 2D versions of the algorithm, these are respectively univariate and bivariate. The algorithm identifies a list of critical articulators for each phone based on the distance of the phone pdfs from the grand distributions. Meanwhile, there are dependent articulators that are influenced by their relation to the critical articulators. So, the pdfs of dependent articulators in the model are adjusted based on the critical articulator pdfs using grand correlation amongst articulators. Kullback-Leibler divergence (KLD) is used as the distance measure between distributions of articulator coordinates (Kullback, 1968), which is minimised during operation as the identification and update steps are repeated.

2.1 Preparation of articulatory data

In this paper, the algorithm was applied to EMA data for two subjects from the MOCHA-TIMIT database (Wrench, 2001). These 14-channel data represent the horizontal (x) and vertical (y) midsagittal movements of 7 fleshpoints: upper lip UL, lower lip LL, lower incisor LI, tongue tip TT, tongue blade TB, tongue dorsum TD and velum V. The movements of these points are calibrated and registered with upper incisor and the bridge of the nose as reference points. Recordings of 460 English TIMIT sentences from one male (msak) and one female (fsew) speaker were used. Their EMA data were smoothed and downsampled to 100 Hz. We use these EMA fleshpoint coordinates as a low-dimensional representation of the articulators. Although they are continuously deformable, a few well-selected points can faithfully represent the full shape of the articulators with reasonable accuracy (Badin and Serrurier, 2006; Qin et al., 2008) For our 1D algorithm, we treat the EMA data as 14 separate 'articulators';

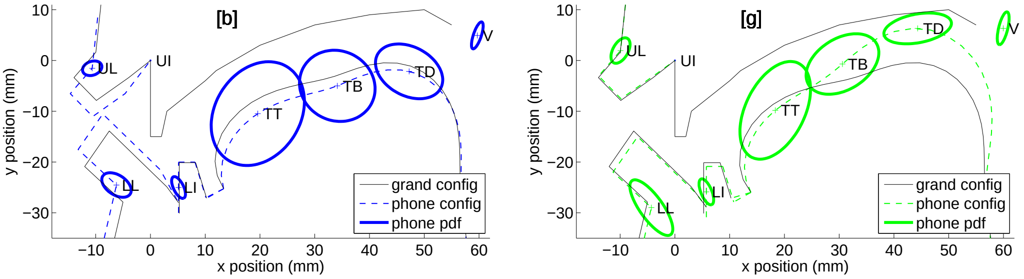


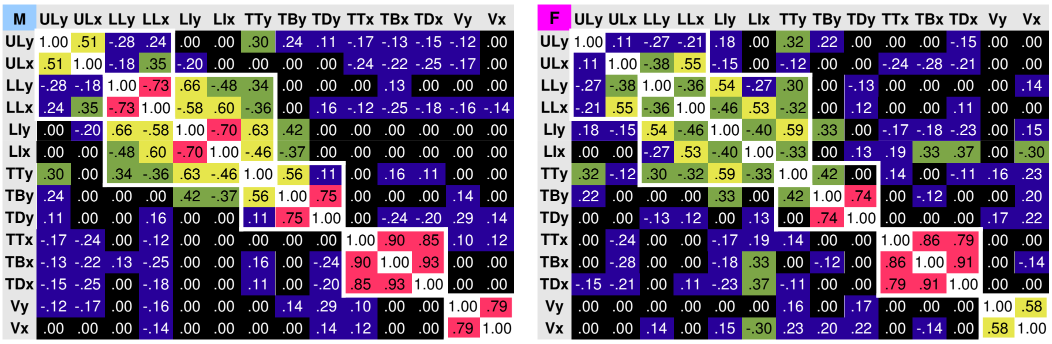
Fig. 2. Outline of grand (thin solid black) and phone (dashed coloured) midsagittal configurations for [b] (left, blue) and [g] (right, green) from male subject, with covariance ellipses (thick solid coloured) representing phone pdfs.

for the 2D case, we combine x and y coordinates to 7 'articulators'.

During our analysis, we discovered some errors with the phone annotations. There were 8 sentences with incorrect text for each subject, which were given phonetic transcriptions derived from the corrected text with manual alignment. One sentence held corrupted EMA data and was discarded. There were some cases where the database's automatic alignment had failed and others where the dictionary transcripts did not match the utterances. These were identified and corrected manually, with particular attention to alveolar obstruents, /t/, /d/ and /n/, which suffered high levels of elision and deletion. Full details of the changes can be found online (Jackson et al., 2004). Samples at phone midpoints were selected to characterise phoneme targets and constitute the phone distributions. From a minimum of 17 for [ʒ] to a maximum ∼1400 for [ə], the average number of samples was 293.

Figure 2 depicts the phone distributions for [b] and [g] overlaid on a schematic outline of the vocal tract.[1] For the bilabial stop [b], distributions of upper and lower lips appear closely constrained whereas the tongue tip, blade and dorsum maintain large variances. For [g], the tongue dorsum, which is critical for producing the velar stop, has a shifted mean and modified covariance compared with the neutral configuration, although there remains considerable variation about the new mean. The tongue blade and tip are affected by the tongue dorsum's movement due to correlations across the tongue, and act as dependent articulators for that phone; distributions of UL and LI show little change and are redundant. The following section describes how these statistics are used to determine the difference articulatory roles.

[1] Vocal tract outlines were derived from mean flesh point positions using splines and other heuristics, for visualisation. The sketched shape of the lips, teeth and velum are linked to the corresponding articulator positions, whereas the hard palate was drawn to circumscribe the complete set of recorded tongue positions.



M	ULy	ULx	LLy	LLx	LIy	LIx	TTy	TBy	TDy	TTx	TBx	TDx	Vy	Vx
ULy	1.00	.51	-.28	.24	.00	.00	.30	.24	.11	-.17	-.13	-.15	-.12	.00
ULx	.51	1.00	-.18	.35	-.20	.00	.00	.00	.00	-.24	-.22	-.25	-.17	.00
LLy	-.28	-.18	1.00	-.73	.66	-.48	.34	.00	.00	.00	.13	.00	.00	.00
LLx	.24	.35	-.73	1.00	-.58	.60	-.36	.00	.16	-.12	-.25	-.18	-.16	-.14
LIy	.00	-.20	.66	-.58	1.00	-.70	.63	.42	.00	.00	.00	.00	.00	.00
LIx	.00	.00	-.48	.60	-.70	1.00	-.46	-.37	.00	.00	.00	.00	.00	.00
TTy	.30	.00	.34	-.36	.63	-.46	1.00	.56	.11	.00	.16	.11	.00	.00
TBy	.24	.00	.00	.00	.42	-.37	.56	1.00	.75	.00	.00	.00	.14	.00
TDy	.11	.00	.00	.16	.00	.00	.11	.75	1.00	.00	-.24	-.20	.29	.14
TTx	-.17	-.24	.00	-.12	.00	.00	.00	.00	.00	1.00	.90	.85	.10	.12
TBx	-.13	-.22	.13	-.25	.00	.00	.16	.00	-.24	.90	1.00	.93	.00	.00
TDx	-.15	-.25	.00	-.18	.00	.00	.11	.00	-.20	.85	.93	1.00	.00	.00
Vy	-.12	-.17	.00	-.16	.00	.00	.00	.14	.29	.10	.00	.00	1.00	.79
Vx	.00	.00	.00	-.14	.00	.00	.00	.00	.14	.12	.00	.00	.79	1.00

F	ULy	ULx	LLy	LLx	LIy	LIx	TTy	TBy	TDy	TTx	TBx	TDx	Vy	Vx
ULy	1.00	.11	-.27	-.21	.18	.00	.32	.22	.00	.00	.00	-.15	.00	.00
ULx	.11	1.00	-.38	.55	-.15	.00	-.12	.00	.00	-.24	-.28	-.21	.00	.00
LLy	-.27	-.38	1.00	-.36	.54	-.27	.30	.00	-.13	.00	.00	.00	.00	.14
LLx	-.21	.55	-.36	1.00	-.46	.53	-.32	.00	.12	.00	.00	.11	.00	.00
LIy	.18	-.15	.54	-.46	1.00	-.40	.59	.33	.00	-.17	-.18	-.23	.00	.15
LIx	.00	.00	-.27	.53	-.40	1.00	-.33	.00	.13	.19	.33	.37	.00	-.30
TTy	.32	-.12	.30	-.32	.59	-.33	1.00	.42	.00	.14	.00	-.11	.16	.23
TBy	.22	.00	.00	.00	.33	.00	.42	1.00	.74	.00	-.12	.00	.00	.20
TDy	.00	.00	-.13	.12	.00	.13	.00	.74	1.00	.00	.00	.00	.17	.22
TTx	.00	-.24	.00	.00	-.17	.19	.14	.00	.00	1.00	.86	.79	.00	.00
TBx	.00	-.28	.00	.00	-.18	.33	.00	-.12	.00	.86	1.00	.91	.00	-.14
TDx	-.15	-.21	.00	.11	-.23	.37	-.11	.00	.00	.79	.91	1.00	.00	.00
Vy	.00	.00	.00	.00	.00	.00	.16	.00	.17	.00	.00	.00	1.00	.58
Vx	.00	.00	.14	.00	.15	-.30	.23	.20	.22	.00	-.14	.00	.58	1.00

Fig. 3. Grand 1D correlation matrix R^* from male (left) and female (right) data containing strong and significant correlations ($|r_{ij}| > 0.1$, $\alpha = 0.05$).

2.2 Proposed algorithm with independent coordinates (1D)

The proposed algorithm identifies critical articulators for each phone based on the distance between grand and phone pdfs, calculated as symmetrical Kullback-Leibler divergence. Using grand correlations amongst articulators, the pdfs of dependent articulators are conditioned on the critical articulator pdfs, reflecting statistical properties of muscle and tissue linkages in speech production. Since some phones engage more than one critical articulator, the algorithm builds up a list of critical (and dependent) articulators incrementally until the model pdfs converge onto the phone pdfs, to within a threshold. Articulators not correlated with any critical articulator are declared redundant. The algorithm was implemented for 1D and 2D pdfs, where x and y coordinates from each flesh point were treated independently (1D), or their covariation incorporated (2D). First, we explain the 1D version of the algorithm.

Derive statistics and initialise model. As a precursor to running the algorithm, it is essential to gather the grand and phone statistics and calculate the significant correlations amongst the articulator coordinates. We look first at the 1D case, which treats x and y coordinates as separate articulatory measurements. Univariate correlations were computed from the 14-channel articulatory data, $R = \{r_{ij}\}$ for $i, j = \{1..a\}$ and a=14. Where there were small or insignificant correlations, we chose to eliminate them to ensure that the parameters we kept were supported by strong statistical evidence. So statistical significance of the correlations was given Pearson's test, and insignificant and weak correlations were set to zero ($\alpha = 0.05$ and $|r_{ij}| < 0.1$). Figure 3 depicts the grand correlation matrix $R^* = \{r^*_{ij}\}$ of the remaining significant and strong correlations for male (left) and female (right) data. So, using M_i and Σ_i denote the grand mean and covariance of each articulator i, the covariance between articulators i and j is $\Sigma_{ij} = \Sigma_i^{1/2} r^*_{ij} \Sigma_j^{1/2}$. The overall mean and covariance for each articulator in the data set were used to define a normally-distributed grand pdf, $\mathcal{N}(M_i, \Sigma_i)$. The subset of data corresponding to each phone ϕ was similarly used to define Gaussian phone pdfs, $\mathcal{N}(\mu_i^\phi, \Sigma_i^\phi)$.

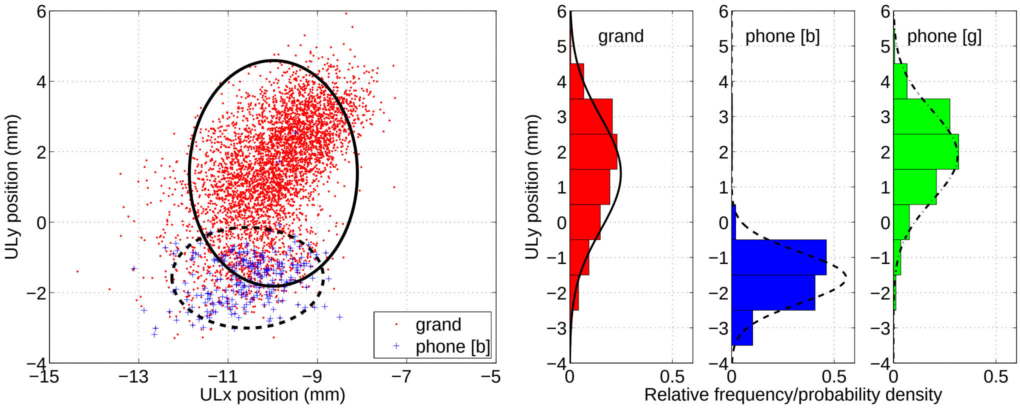


Fig. 4. Left: scatter plot of grand (red dot) and phone [b] (blue plus) upper lip coordinates (ULx and ULy) at the centre of each phone label for male subject's data; ellipses show ±2 standard deviations of the corresponding grand (solid) and phone (dashed) normal distributions, which encompass 95% of the points. Right: histograms with fitted Gaussian distributions of the vertical upper lip coordinate (ULy) for grand (red with solid black), phone [b] (blue with dashed black) and phone [g] (green with dash-dot black).

Highlighting the biomechanical behaviour of human speech apparatus, Figure 3 (left) shows the strong and significant correlations between tongue tip TT, blade TB and dorsum TD in the x direction and the y direction, although there was almost no correlation between the tongue's x and y coordinates. The y correlations (vertical or transverse) were less than in the x direction (horizontal or axial). The absence of correlation between TTy and TDy also shows how the vertical movement of the tongue tip was independent from that of the tongue dorsum. The lower lip LL was strongly correlated in both x and y directions to the jaw LI and, to a lesser extent, to the upper lip UL. Velum x and y movements showed strong correlation, but little with other articulators. Some correlation existed between the jaw LIy and TTy, but otherwise the articulatory system behaved like three largely-independent components: the lip and jaw group, the tongue and the velum. The main patterns for the female speaker, in Figure 3 (right), were consistent with the male correlations. Some differences were possibly due to speaking style and anatomy: correlations between x and y movements of UL, LL and LI were weaker; ULy with LLx had negative correlation; LIy with TTx, TBx and TDx were stronger. Across both speakers, slightly less than half of the correlations (37%) were found to be statistically insignificant or weak. At the start of the algorithm's operation, we define model distributions for the current phone and set them equal to the grand pdfs: $\mathcal{N}(m_i^{\phi,0}, S_i^{\phi,0}) = \mathcal{N}(M_i, \Sigma_i)$.

Compute KLD. The KLD, which provides the distance metric between model and phone distributions, is based on the integral of their log-likelihood ratios and is related to the Bayes factor. Unlike other measures based on the difference between grand and phone means (e.g., Mahalanobis distance, student's t-test, Hotelling's T^2 test) or ratio of variances (Fischer's linear discriminant), the KLD incorporates changes both in mean and covariance.

The symmetric KLD is calculated for each articulator as (Kullback, 1968):

$$
\begin{aligned}
J &= \int_{-\infty}^{\infty} f_1(x) \ln \frac{f_1(x)}{f_2(x)}\, dx \;+\; \int_{-\infty}^{\infty} f_2(x) \ln \frac{f_2(x)}{f_1(x)}\, dx \\
&= \frac{1}{2} \left[\mathrm{tr}\,(S - \Sigma) \left(\Sigma^{-1} - S^{-1} \right) + \mathrm{tr} \left(S^{-1} + \Sigma^{-1} \right) (m - \mu)\,(m - \mu)^{\mathsf{T}} \right] \quad (1)
\end{aligned}
$$

where tr$(\cdot)$ denotes the trace of the matrix, $^{\mathsf{T}}$ denotes the transpose, and the pdfs that we consider are assumed to be Gaussian, $f_1 = \mathcal{N}(m, S)$ and $f_2 = \mathcal{N}(\mu, \Sigma)$. The KLD values lie between zero (for perfectly matching distributions) and infinity. We made maximum likelihood estimates of the Gaussian distribution of phone data, using the sample mean and variance. Next, we took account of the uncertainty over the precise position of the distribution by including the estimation error of the mean (i.e., the standard error $\sqrt{\Sigma/N}$ with N samples) as an additional source of variation. For example, the variance Σ we used in the KLD was replaced by $\Sigma + \Sigma/N$, which allowed us to factor in the range of all likely distributions within one calculation. This is a conservative step, which helps to regularise the small-sample distributions for the KLD calculation, although it makes little difference to the empirical results.

Figure 4 illustrates distributions used in the KLD calculation for two phones, [b] and [g]. Fig. 4 (left) gives a scatter plot of UL data for phone [b] and overall. Fig. 4 (right) shows the distribution of ULy samples overall (grand), and for [b] and [g]. The KLD between grand and phone [b] was high (9.1), identifying ULy as critical for the labial. As a consequence, the correlated LLy is flagged as dependent. Whereas for [g], the divergence was low (0.2) and ULy therefore redundant for the velar.

C step: identify next critical articulator. The algorithm iteratively identifies a list of critical articulators for each phone, initialising a model with the grand pdfs and converging towards the phone pdfs. In the C step, the KLD between the model and phone distributions, termed the *identification divergence*, was the distance metric used to identify as critical the articulator whose pdf differed most from the phone pdf.

The algorithm detail is given in Figure 5, for which we define the phone set Φ, the grand statistics as R^* and $\Gamma = \{M, \Sigma, N\}$ where N is the total sample size, and the phone statistics as the inter-articulator correlations R^ϕ and $\Lambda^\phi = \{\mu^\phi, \Sigma^\phi, \nu^\phi\}$, where ν^ϕ denotes the sample size for phone ϕ. The algorithm iterates through the levels, $k = 0..a$, which determine the length of critical articulator list $C^{\phi,k}$. The function *computeIdiv* computes the identification divergence, $J_i^{\phi,k}$, and the state of the model is denoted by $\Delta^{\phi,k} = \{m^{\phi,k}, S^{\phi,k}, n^{\phi,k}\}$ which consists of the set of model means $m_i^{\phi,k}$, variances $S_i^{\phi,k}$ and sample sizes $n_i^{\phi,k}$ for each articulator i. At each level k, the algorithm considers $(a-k)$ candidates to extend the list to the next level. The algorithm selects the articulator j with the maximum divergence. If j's divergence exceeds the convergence threshold θ_{C}, it is added to the critical list and its distribution set equal to the phone pdf, $\mathcal{N}\left(m_j^{\phi,k}, S_j^{\phi,k}\right) = \mathcal{N}\left(\mu_j^{\phi}, \Sigma_j^{\phi}\right)$.

D step: update dependent articulators. In the D step, a dependent articulator i (one substantially correlated with the critical articulator) has its pdf, $\mathcal{N}\left(m_i^{\phi,k}, S_i^{\phi,k}\right)$, adjusted according to the effect of specifying the critical articulator (Anderson, 1984):

$$m_i = M_i + \Sigma_{ij}\Sigma_j^{-1}\left(m_j - M_j\right) \tag{2}$$

$$S_i = \Sigma_i + \Sigma_{ij}\Sigma_j^{-1}\left(S_j - \Sigma_j\right)\Sigma_j^{-1}\Sigma_{ji}. \tag{3}$$

The distributions of dependent articulators are updated based on all critical articulators identified up to level k, using the grand and phone statistics, Γ and Λ^{ϕ}. The dependence threshold θ_{D} avoids over-updating the dependent articulators. A value $\theta_{\mathrm{D}} = 0.1$ was used for all experiments. The algorithm repeats the KLD calculation, the C step and the D step, until all $J^{\phi,k}$ are within tolerance θ_{C}. Thus, we obtain the final list of critical, dependent and (by elimination) redundant articulators for phone ϕ. The procedure is repeated for all phones in the inventory, $\phi \in \Phi$.

The upper part of Figure 6 illustrates operation of the 1D algorithm on male speaker data for [b]. Grand, phone and model distributions are represented by dotted-red, dashed-green and solid-blue variance ellipses respectively, which show ±2 standard deviations around the mean. The upper lip was identified as the first critical articulator (Fig. 6, left), having maximum identification divergence of the initial model (grand) pdf from the phone pdf. The 1D model distribution for ULy was set to the phone pdf in the algorithm's C step. The effect of the lip's configuration on the other articulators was calculated in the D step from the grand correlations (Fig. 6, top centre): stronger correlations induced greater effects. Here, ULy influenced ULx and LLy, but there was no change in the LIx or LIy distributions, for instance, since they had no substantial correlation with ULy. On the other hand, the horizontal tongue coordinates TTx, TBx and TDx were not updated because their distributions already matched the corresponding phone pdfs, giving divergences below the dependence threshold θ_{D}. At the next level (Fig. 6, top right), the lower lip height LLy was chosen as critical, which adjusted the jaw distribution LIy, and LLx taking both critical articulators so far identified into account. With all identification divergences below the convergence threshold $\theta_{\mathrm{C}} = 1.7$, no

```
**Derive statistics**
Global statistics $\mathbf{\Gamma} = \{\mathbf{M}, \mathbf{\Sigma}, N\}$: means ($a \times K$), variances ($a \times K \times K$) and sample size, and correlation $\mathbf{R}^*$
Phone statistics $\mathbf{\Lambda}^\phi = \{\boldsymbol{\mu}^\phi, \mathbf{\Sigma}^\phi, \nu^\phi\}$: means ($a \times K$), variances ($a \times K \times K$) and sample size, and correlation $\mathbf{R}^\phi$
Model statistics $\mathbf{\Delta}^{\phi,k} = \{\mathbf{m}^{\phi,k}, \mathbf{S}^{\phi,k}, n^{\phi,k}\}$: means ($a \times K$), variances ($a \times K \times K$) and sample size ($a \times 1$)
Threshold $\Theta = \{\theta_\mathrm{C}, \theta_\mathrm{D}\}$
**Initialise model**
$\mathbf{m}_i^{\phi,0} = \mathbf{M}_i$; $\mathbf{S}_i^{\phi,0} = \mathbf{\Sigma}_i$; $n_i^{\phi,0} = N$, for all articulators $i = \{1..a\}$
Empty critical articulator list: $\mathbf{C}^{\phi,0} = \{\,\}$
Prepare for main loop: $k = 0$; $isConverged =$ FALSE
WHILE (($k \leq a$) AND (!$isConverged$))
        **Compute identification divergence**
        $J_i^{\phi,k} = computeIdiv(\mathbf{\Delta}_i^{\phi,k}, \mathbf{\Lambda}_i^\phi)$, for all articulators $i = \{1..a\}$
        Find articulator with maximum divergence: $j = \mathrm{argmax}\{J_1^{\phi,k} .. J_a^{\phi,k}\}$
        **C step**
        IF ($J_j^{\phi,k} > \theta_\mathrm{C}$)
                Increment the level: $k \mapsfrom k + 1$
                Replicate model: $\mathbf{\Delta}^{\phi,k} = \mathbf{\Delta}^{\phi,k-1}$
                Add critical articulator: $\mathbf{C}^{\phi,k} = \{\mathbf{C}^{\phi,k-1}\} \cup \{j\}$
                Set distribution: $\mathbf{m}_j^{\phi,k} \mapsfrom \boldsymbol{\mu}_j^\phi$;  $\mathbf{S}_j^{\phi,k} \mapsfrom \mathbf{\Sigma}_j^\phi$;  $n_j^{\phi,k} \mapsfrom \nu^\phi$
                **D step**
                $\mathbf{\Delta}^{\phi,k} = updateDep(\mathbf{\Gamma}, \mathbf{R}^*, \mathbf{\Lambda}^\phi, \mathbf{R}^\phi, \mathbf{\Delta}^{\phi,k}, \Theta, J^{\phi,k-1}, \mathbf{C}^{\phi,k})$
        ELSE
                $isConverged =$ TRUE
                Store final critical articulator list: $\hat{\mathbf{C}}^\phi = \mathbf{C}^{\phi,k}$
                Store model statistics: $\hat{\mathbf{m}}^\phi = \mathbf{m}^{\phi,k}$;  $\hat{\mathbf{S}}^\phi = \mathbf{S}^{\phi,k}$;  $\hat{n}^\phi = n^{\phi,k}$
                Store no. critical articulators: $\hat{k}^\phi = k$
        END IF
END WHILE

**function computeIdiv**$(\mathbf{\Delta}_i^{\phi,k}, \mathbf{\Lambda}_i^\phi)$
        Incorporate standard error: $\mathbf{S}_1 = \mathbf{S}_i^{\phi,k} + (\mathbf{S}_i^{\phi,k}/n_i^{\phi,k})$;  $\mathbf{S}_2 = \mathbf{\Sigma}_i^\phi + (\mathbf{\Sigma}_i^\phi/\nu^\phi)$
        $J = \frac{1}{2}\left(\mathrm{tr}(\mathbf{S}_1 - \mathbf{S}_2)(\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1}) + \mathrm{tr}(\mathbf{S}_1^{-1} + \mathbf{S}_2^{-1})(\mathbf{m}_i^{\phi,k} - \boldsymbol{\mu}_i^\phi)(\mathbf{m}_i^{\phi,k} - \boldsymbol{\mu}_i^\phi)^\mathsf{T}\right)$
RETURN $J$

**function updateDep**$(\mathbf{\Gamma}, \mathbf{R}^*, \mathbf{\Lambda}^\phi, \mathbf{R}^\phi, \mathbf{\Delta}^{\phi,k}, \Theta, J^{\phi,k-1}, \mathbf{C}^{\phi,k})$
        Get critical grand statistics from $\mathbf{\Gamma}$ and $\mathbf{R}^*$: $\mathbf{M}_\mathbf{C} = \{\mathbf{M}_i\}_{i \in \mathbf{C}^{\phi,k}}$;  $\mathbf{\Sigma}_\mathbf{CC} = \{\mathbf{\Sigma}_{ij}\}_{i,j \in \mathbf{C}^{\phi,k}}$
        Get critical phone statistics from $\mathbf{\Lambda}^\phi$ and $\mathbf{R}^\phi$: $\boldsymbol{\mu}_\mathbf{C}^\phi = \{\boldsymbol{\mu}_i^\phi\}_{i \in \mathbf{C}^{\phi,k}}$;  $\mathbf{\Sigma}_\mathbf{CC}^\phi = \{\mathbf{\Sigma}_{ij}^\phi\}_{i,j \in \mathbf{C}^{\phi,k}}$
        FOR $i = \{1..a\} - \{\mathbf{C}^{\phi,k}\}$
                IF ($J_i^{\phi,k-1} > \theta_D$)
                        Get dependent covariance: $\mathbf{\Sigma}_{i\mathbf{C}} = \{\mathbf{\Sigma}_{ij}\}_{j \in \mathbf{C}^{\phi,k}}$
                        Update mean: $\mathbf{m}_i^{\phi,k} \mapsfrom \mathbf{M}_i + \mathbf{\Sigma}_{i\mathbf{C}}\mathbf{\Sigma}_\mathbf{CC}^{-1}\left(\boldsymbol{\mu}_\mathbf{C}^\phi - \mathbf{M}_\mathbf{C}\right)$
                        Update variance: $\mathbf{S}_i^{\phi,k} \mapsfrom \mathbf{\Sigma}_i + \mathbf{\Sigma}_{i\mathbf{C}}\mathbf{\Sigma}_\mathbf{CC}^{-1}\left(\mathbf{\Sigma}_\mathbf{CC}^\phi - \mathbf{\Sigma}_\mathbf{CC}\right)\mathbf{\Sigma}_\mathbf{CC}^{-1}\mathbf{\Sigma}_{i\mathbf{C}}^\mathsf{T}$
                        Update sample size: $n_i^{\phi,k} \mapsfrom \nu^\phi$
                END IF
        END FOR
RETURN $\mathbf{\Delta}^{\phi,k}$
```

Fig. 5. Algorithm for articulatory role identification for phone ϕ, including functions for computnig KLD and updating model distributions using critical articulator information and inter-articulator correlations. The 1D (K=1) or 2D (K=2) versions use scalar or vector means, $\boldsymbol{M}$, $\boldsymbol{\mu}^\phi$ and $\boldsymbol{m}^{\phi,k}$, and scalar or matrix (co-)variances, $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^\phi$ and $\boldsymbol{S}^{\phi,k}$.

further articulators were identified here. This demonstrates how the correlations of dependent articulators affect the model distributions that are used for identifying subsequent critical articulators. Given ULy and LLy were critical, the grand correlations offered some dependence for all other articulators except the velum Vx. So, the dependent articulators were ULx, LLx, LIy, TTy,

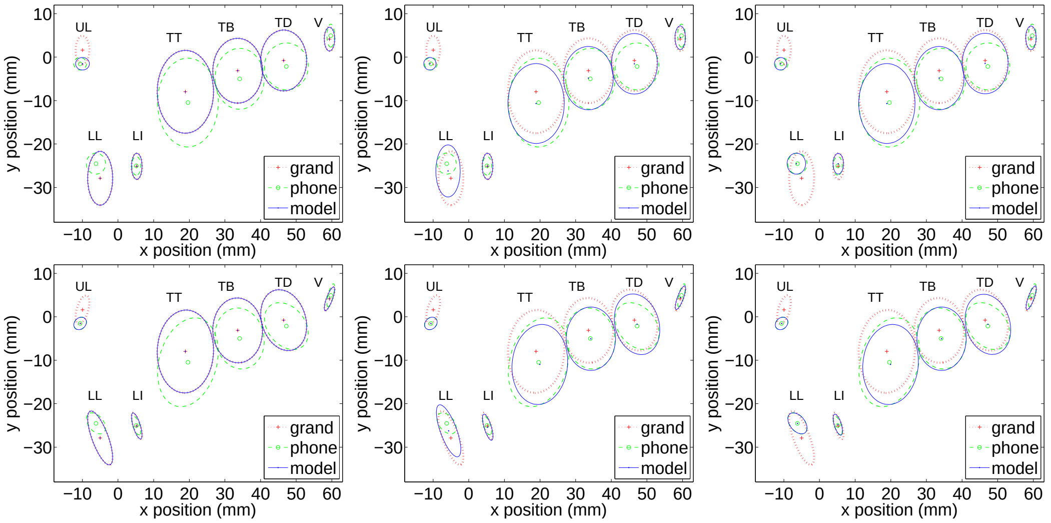


Fig. 6. Mid-sagittal schematic of convergence of 1D (upper) and 2D (lower) phone models of [b] for male data, using grand (thick dotted red, +), phone (medium dashed green, ∘) and model (thin solid blue, ·) distributions as the algorithm progresses (from left to right): first critical articulator is identified k=1, dependent articulators are updated k=1, second iteration is completed k=2.

TBy, TDy, and Vy. This is largely expected as a consequence of jaw raising, although the small tongue and velum correlations are all less than one third. The redundant articulators were LIx, TTx, TBx, TDx, and Vx.

2.3 Proposed algorithm with correlated coordinates (2D)

The major and minor axes of the drawn ellipses are aligned with the covariance's eigenvectors. The 2D version of the algorithm follows a similar course to the 1D version, selecting UL then LL in this case. It incorporates the covariation found within an articulator's x and y movements, and updates dependent articulators in line with canonical correlations, as now described.

The lower part of Figure 6 shows the algorithm working with 2D distributions, where bivariate correlations from the 2D articulatory data are computed using canonical correlation analysis (Johnson and Wichern, 1998), which finds orthogonal directions in which a pair of articulators, i and j, are maximally correlated: $\boldsymbol{\rho}_{ij} = \mathrm{diag}(\rho^1_{ij}, \rho^2_{ij})$ denotes the canonical correlations, and $\mathbf{U}_i$ and $\mathbf{U}_j$ the corresponding 2D eigenvectors. Like before, statistically insignificant and weak canonical correlation values are set to zero ($\alpha = 0.05$, $|\rho_{ij}| < 0.15$). This gives the bivariate correlation matrix $\mathbf{R}^* = \{\mathbf{r}^*_{ij}\}$, for $i, j = 1..7$, where $\mathbf{r}^*_{ij} = \mathbf{U}_i \boldsymbol{\rho}_{ij} \mathbf{U}^\mathsf{T}_j$. The covariance between i and j is $\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}^{1/2}_i \mathbf{r}^*_{ij} \boldsymbol{\Sigma}^{1/2}_j$, where $\boldsymbol{M}_i$ and $\boldsymbol{\Sigma}_i$ denote the grand 2D mean vector and covariance matrix of articulator i.

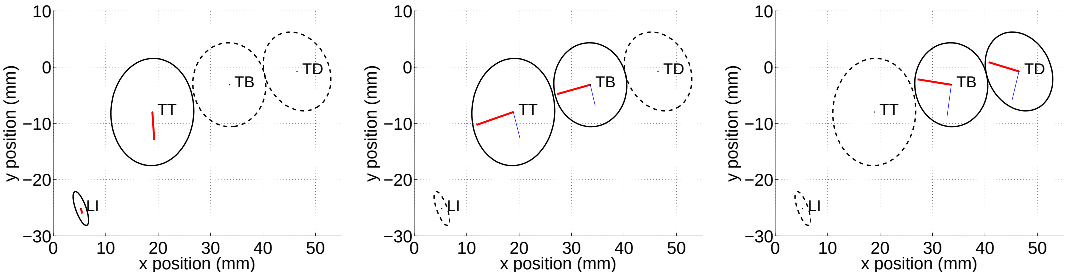


Fig. 7. First (thick red) and second (thin blue) canonical correlations for male data, within their respective (solid) grand covariance ellipses (from left): LI-TT, TT-TB and TB-TD.

To illustrate the effect of this transformation, Figure 7 shows the direction and strength of canonical covariation between three adjacent fleshpoint pairs (male, from left): LI-TT, TT-TB and TB-TD. Only one significant canonical correlation from jaw rotation was found for the first pair: $\rho^1_{\mathrm{LI,TT}} = 0.62, \rho^2_{\mathrm{LI,TT}} = 0$. For the last two pairs along the tongue, the eigenvector directions indicate primary correlation for axial (forward/backward) motion ($\rho^1_{\mathrm{TT,TB}} = 0.93, \rho^1_{\mathrm{TB,TD}} = 0.93$), and secondary for raising/lowering ($\rho^2_{\mathrm{TT,TB}} = 0.53, \rho^2_{\mathrm{TB,TD}} = 0.75$). A third of the articulator combinations had two significant, strong correlations, no correlation was found for one pair in seven, and the rest (52%) had one significant canonical correlation. The female data painted a very similar picture (2 sig. ×8 pairs, 1 sig. ×11 pairs, 0 sig. ×2 pairs), and the bivariate correlations were similar to the univariate correlations in absolute value. Although the 1D correlations capture the relation between critical articulator dimensions and those of dependent articulators, the 2D canonical correlations provide additional accuracy by modeling the correlations within each critical and dependent articulator.

In the 2D version of the algorithm, the grand, model and phone distributions are assumed to be bivariate Gaussians with a=7. The identification divergence is thus defined as the KLD between 2D model and phone distributions, where the means, $\boldsymbol{m}$ and $\boldsymbol{\mu}$, in eq. 1 are 2D column vectors and the variances, $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$, are 2×2 matrices.

3 Running the algorithm with EMA data

This section assesses the factors limiting the performance of the proposed method on the MOCHA-TIMIT articulatory data.

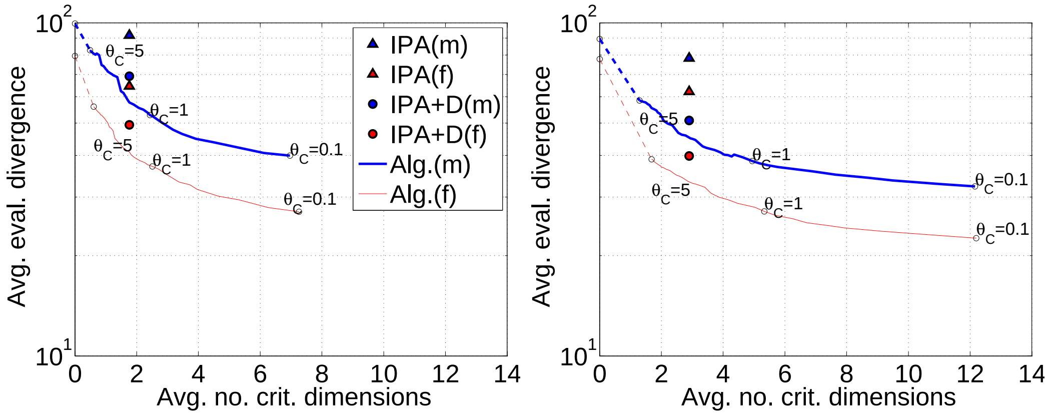


Fig. 8. Convergence of 1D (left) and 2D (right) algorithms: evaluation divergence between 14D phone pdfs and the model pdfs with critical threshold $\theta_\mathrm{C} = \{0.1, 0.2, \ldots, 5\}$, averaged over Φ; male (thick blue line), female (thin red line). Symbols show divergence between phone pdfs and ones based on IPA-derived phonetic rules fitted the data, without (△) and with (○) considering articulator dependencies, for male (blue/filled) and female (red/open) data.

3.1 Assessment of model convergence

As we saw in Figure 6, the model pdfs tend toward those of the phone as the critical articulators are identified. To evaluate the goodness of fit of model pdfs to the phone distributions, we defined an *evaluation divergence* as the full dimensional KL divergence between the model and phone distributions. The 1D model means and variances of each articulatory dimension were collated into a 14D mean vector and 14D diagonal covariance matrix. With the 2D models, the intra-articulator x and y covariation was included in the covariance matrix either side of the diagonal. The covariance matrix of each phone distribution was the full 14D, whereas the model covariance matrices reflected the 1D and 2D assumptions. The divergence was calculated as in equation (1) using function *computeIdiv* in Fig. 5. The convergence of the model was evaluated as a function of the convergence threshold θ_C, which is applied to the 1D and 2D identification divergence. As we reduce the threshold, we expect the algorithm to identify a greater number of critical articulators and to match the phone pdfs more closely. This evaluation procedure allows us to quantify the goodness of the fit, to test how it trades with θ_C, and to compare the 1D and 2D versions of the algorithm.

3.2 Effect of the convergence threshold

Using the evaluation divergence, we measured the effect of adjusting the convergence threshold θ_C. Figure 8 shows the evaluation divergence between the 14D phone pdfs and the collated 1D and 2D models. Each plotted point marks one value of the threshold θ_C, which was reduced from 5.0 to 0.1 in steps of 0.1. The horizontal axis reports the average number of critical dimensions per phone (number of critical articulators times model dimensionality), and the vertical axis shows the mean evaluation divergence over the phone set Φ. The mean divergence between the initial 1D model (diagonal covariance) and the full covariance phone pdfs was 99 for the male speaker and 80 for the female, who was consistently lower at all θ_C values. As the threshold was lowered (Fig. 8, left), convergence improved at the expense of increased critical dimensions: a 32% reduction in the male evaluation divergence was achieved and critical dimensions rose from 0.4 to 3, as the threshold went from 5 to 1. A similar trend was observed in female divergence. Lowering the threshold down to 0.1 claimed half of the articulatory dimensions (∼7) and reduced divergence by a further 15% for both speakers.

The fit of the 2D model pdfs to the full covariance phone models includes correlations between x and y movements within each measured articulatory fleshpoint (Fig. 8, right). The initial divergence was cut by 12% and 3% for male and female data relative to the initial 1D evaluation divergence. The trend was for similar reductions (25% male, 20% female) as θ_{C} was lowered from 5 to 1, increasing critical dimensions per phone from 2 to 6. Convergence improved by a mere 5% for $\theta_C = 0.1$ with all but one articulator on average deemed critical (∼12 critical dimensions), for both male and female speakers. With the additional flexibility to describe correlations between x and y movements of each articulator, the fit of the 2D model pdfs to the (full covariance) phone pdfs was better than the 1D pdfs, at all levels of threshold.

3.3 Limitations of the data set

The MOCHA-TIMIT database has provided a valuable number of EMA recordings over many read sentences and two labeled subjects. Having corrected some alignment and transcription errors in the phone labels, the mid-phone sampled distributions gave sufficient information to determine good lists of critical, dependent and redundant articulators, despite known problems with fleshpoint calibration (Frankel, 2003; Richmond, 2007). Improvements could be made, therefore, with further recordings of speech articulation, e.g., new capture techniques, larger corpora, multiple subjects and various speaking styles. The algorithm can be readily extended to work with 3D data. Also, the model pdfs in the algorithm could be developed better to fit measured distributions, e.g., by Gaussian mixture or numerical methods (Hershey and Olsen, 2007).

4 Analysis of phonological findings

The previous section's quantitative analysis shows the trade off between modeling accuracy and model complexity for two different versions, 1D and 2D. It also demonstrates that the algorithm captures the characteristics of the phone distributions more effectively than phonetic rules derived from a conventional IPA description at an equivalent level of complexity, which will now be described in detail. To interpret the results in the context of IPA's place and manner of articulation, we compare of the automatically-identified critical articulators for each phone with the phonetic descriptions from the IPA chart.

4.1 Comparison with IPA

The lists of critical articulators that were obtained for each phone are compared with the international phonetic alphabet (IPA) descriptions for consonants and vowels. IPA is a widely accepted representation of a language's phonetic repertory that provides an articulatory description of speech sounds, which lets us make a comparative analysis of the results obtained using our role identification algorithm. Although the IPA has been designed as a notational standard for phonetic description of the speech sound categories in the world's languages, it provides a good basis for comparison because it is an internationally agreed summary of the knowledge built up over many generations. Its main purpose involves the transcription of human speech by phoneticians, and is therefore tailored (i) to encapsulate meaningful distinctions in the context of language, (ii) for utterances produced by humans and (iii) observed by phoneticians. Increasingly, there is a need for phonological descriptions for use in speech technologies that can (i) model the characteristics of typical phones within a language, (ii) include the implicit effects found in human phoneme-to-phone realization, such as coarticulation, and (iii) incorporate knowledge from other types of observations, such as spectrograms, X-ray, MRI and articulography data. For purposes of comparison with the results of the proposed algorithm, IPA-based pdfs were generated by using IPA manner and place attributes to define a set of critical articulators for each phone, in contrast to the algorithm's C step. For both 1D and 2D versions, two kinds of realization were considered: one that only updated the IPA-specified articulators, and one that also incorporated their effect on dependent articulators. As before, the evaluation divergence measured the goodness of fit to the recorded phone distribution data.

Simply setting the distributions equal to the phone pdfs for the articulators specified by IPA (the remainder equal to the grand pdfs) gave little benefit (Fig. 8): the evaluation divergence was similar to that between the phone pdfs and the grand pdfs. For IPA-based consonant pdfs, including articulator dependencies via the algorithm's D-step made divergence fall by approximately 30% (1D: 24% male and 23% female; 2D: 35% male and 37% female). However, at the same number of average critical dimensions per phone, the model pdfs using articulators identified by our proposed algorithm showed reductions of approximately 40% (1D: 37%/37%; 2D: 43%/45% for male/female). So, given that there is little dispute over the correct classification of most phonemes' manner and place, what differences enable the algorithm to obtain a better fit to the observed data without increasing the level of complexity? The answer, below, lies in the detail of the critical articulators specified for each phone.

4.2 Consonants

Place and manner descriptors from the IPA chart were used to specify a set of critical articulators for each phone, for both 1D and 2D cases. Beginning with the consonants, place descriptors were related to the articulators actively involved in the realization of phonemes, as follows for 1D: ULy and LLy for bilabials /p,b,m/; LLx and LLy for labio-dentals /f,v/; TTx and TTy for inter-dentals, alveolars and post-alveolars /θ,ð/, /s,z,l,ɹ/ and /ʃ,ʒ,ʧ,ʤ/; TBx and TBy for palatal /j/; only TTy for alveolar obstruents /t,d,n/ and TDy for velars /k,g,ŋ/. For nasals /m,n,ŋ/, Vx was also made critical. For labio-velar /w/, ULx, LLx and TDy were specified. None was specified for glottal /h/ from the available articulator measurement points. Thus, the average was 1.8 critical articulator coordinates or dimensions. Similar lists were specified in the 2D case, yielding an average 2.8 articulatory dimensions or 1.4 articulators per phone.

For fair comparison, the convergence threshold θ_C was set to give the same number of critical articulators as the IPA descriptions (1D: $\theta_C = 1.7$ for both speakers; 2D: $\theta_C = 2.3/2.0$ for male/female). Binary features were obtained to describe the direction of articulator shifts from the neutral (grand mean) position. The statistical significance of any shift was determined using a student's t-test ($\alpha = 0.05$, with compensation for non-homogeneous grand and model variances). At the chosen thresholds, the proposed algorithm identified critical articulators for over 90% of the consonants, and the patterns of results were similar for 1D and 2D. For the glottal [h], no articulators were identified as critical. No critical articulators were identified for the lateral alveolar [l] from the mid-sagittal EMA data. The critical articulators that were identified generally agreed with those derived from the IPA chart, but there were some notable differences.

The velum Vx was identified as critical for the male nasals at the chosen threshold; for the female, the velum had next highest KLD, albeit below the threshold. For both speakers, binary positional features consistently showed the velum displacement for all nasal and oral (non-nasal) sounds. Setting a lower threshold identified one velum dimension as critical for nasal sounds.

For all sibilants [s,z,ʃ,ʒ] and affricates [ʧ,ʤ], the vertical tongue tip movement TTy (normally considered to be the primary articulation) was identified as critical. Yet, even considering their interdependency, the jaw position LIy was equally identified as a critical articulator. The position of the lower teeth may thus be considered a secondary articulation for these sounds, for which the tongue tip constricts the flow into a jet, creating turbulence that impinges on the teeth. This mechanism hugely increases the acoustic efficiency of the sibilant noise source, so it is important to ensure the jet and obstacle are aligned (Shadle, 1985).

Some differences between the expected and identified critical articulators were inconsistent across speakers. For [θ] (female), the tongue blade TBy also lowered to achieve the expected tip position; there was no significant change in the position of the male tongue blade which was not identified as critical. For [ʃ,ʒ,ʤ], the positions of either TB or TD were different for male and female speakers, leading to an extra critical tongue dimension. The number of critical dimensions was the highest for [ʒ], despite standard-error compensation for the small sample size, although its voiceless counterpart [ʃ] and the affricates were equal second in terms of critical articulator list length [ʧ,ʤ]. This finding suggests that fricatives and affricates are the most constrained phonemes to pronounce in British English.

The stops showed the best agreement with the expected set of critical articulators (75%). Only the upper lip was chosen for [p] (both) and TTy for alveolars [t,d]. This echoes the constraint on TDy but not TDx for velars. For two or three phones, the algorithm identified articulators that were strongly correlated with those anticipated. For [t] (female), LIy was selected in addition to TTy. For [w] (male and female), a vertical upper lip gesture ULy was identified which was highly correlated with the expected horizontal specification of ULx and LLx. For [ɹ] (male), the expected and identified critical articulators differed by one tongue position: TB replaced TT. Only TBy was identified for [y].

Having defined place and manner descriptions of consonants from the IPA chart in terms of articulatory coordinates, the algorithm's convergence threshold was adjusted to match the number of articulatory constraints identified. No critical articulation was specified for [h] (male) and the alveolar [l] (both). Sibilant fricatives identified a secondary articulation, of the lower incisors, Few substitutions and insertions were made of correlated articulators. Some male-female differences were found in the tongue position, which could be attributed to individual anatomical or stylistic variation until further data are available. The findings demonstrate that the algorithm not only produces plausible explanations that fit the observations, showing broad agreement with the IPA descriptions, but has capability to make explicit significant details that might be implicit or absent from IPA alone. Hence, we conclude that the algorithm can be used to determine from data the constraints that are important in speech articulation.

4.3 Vowels

Unlike consonants, vowel phonemes do not have such well-defined places of articulation. The target articulatory configurations shown on the IPA chart are part acoustic and part articulatory, so it is not trivial to state the critical articulators for vowels. Tongue height, the backness of the tongue and lip rounding are the key factors that describe vowels' articulatory configurations. Thus, we might expect the tongue y-dimension to be critical for closed (high) vowels, the jaw for open (low) vowels, tongue x for front and back configurations, and lip x for rounded ones. The point on the tongue chosen as critical would likely reflect backness: i.e., TT for front [æ, ɛ, ɪ, iː, i], TB for mid [ə, ɚ, ʌ] and TD for back [ɑ, ɒ, ɔ, ʊ, u] vowels.

Using the same thresholds as for the consonants, the algorithm identified critical articulators for three quarters of the vowels. No critical articulators were identified for [ə, ɪ, ʊ] for either speaker in 1D or 2D cases, which reflects schwa's general properties across the phonetic inventory (rather than a precise neutral configuration) and weak distinctions for the near front and back vowels [ɪ] and [ʊ]. No critical articulators were identified for the open mid back vowel [u] with male data.

Though various parts of the tongue are involved in shaping the vocal tract, our algorithm showed the tongue blade and dorsum to be most critical for production of vowels, together with the lower lip. For close front vowels [iː, i], TTx was identified once for male and female; whereas TBx or TDx were identified two or three times for open back vowels [ɔ, ɒ, ɑ]. In the vertical direction, high front and back vowels selected TBy or TDy; open and low vowels tended to choose LLy. Low back vowels also picked TBy.

The IPA specification of vowel height, backness and roundedness was described in terms of articulatory dimensions. The tongue blade and tongue dorsum featured strongly amongst most vowels, but no critical articulation was specified for 1 in 4 vowels: the short, central and reduced ones, including schwa. Lower lip caught the open vowels and tongue dorsum the closed ones. Lip rounding was not clearly evidenced in the mid-sagittal data. Although less straightforward to specify, the critical articulations identified for vowels reflect the main tendencies across the vowel quadrilateral, and indicate the different nature of

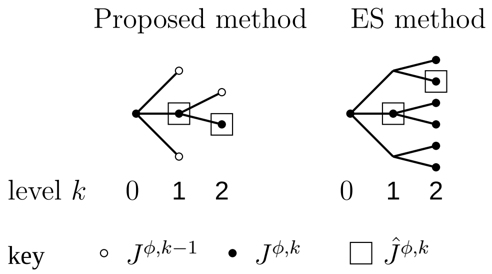


Fig. 9. Proposed depth-first (left) and exhaustive (right) searches for best critical articulator combination.

their constraints, especially for short and mid vowels. These results of automatic identification offer some interesting statistical insights into real-world production of phones that are largely compatible with the IPA description framework. While IPA gives a constant level of specificity to all phones (approximately, at least within broad categories), the algorithm is able to identify pronunication constraints from the data, as they are needed.

5 Algorithm evaluation

5.1 Comparison with exhaustive search

The proposed algorithm performs a kind of greedy or depth-first search (DFS) in identifying a new critical articulator at each level k, depending on the previously identified critical articulators up to level $(k-1)$. To evaluate the performance of the algorithm, an exhaustive search (ES) procedure was adopted that, at each level, searches through all possible combinations of the critical articulators. The ES proceeds independent of any previously selected set of critical articulators, and uses a minimax approach to identify the best critical articulator combination (Coppin, 2004). Thus, the ES finds the globally optimal combination of critical articulators to match model and phone pdfs, for a given level of model complexity. The ES and proposed DFS algorithms' performance was compared using the evaluation divergence, and their critical articulator lists analysed level by level: (i) to see whether the identified critical articulators for each phone differ, (ii) to determine any effect on model convergence of changes in the critical articulator order, and (iii) to find which search procedure gives better fitting models.

5.2 Exhaustive search method

The model statistics were initialized with the grand distributions, as before. However for the exhaustive search, all possible articulator combinations were considered at each subsequent level, $k = 1..a$, which number P^k. This is depicted in Figure 9. For each phone ϕ, the identification divergences between the model and phone pdfs $J_i^{\phi,k}$ were computed by the *computeIdiv* function for all P^k combinations at level k. According to the minimax criterion, the articulator combination that minimized the maximum divergence was chosen as the critical set at that level, $\hat{C}^{\phi,k}$. The number of articulatory combinations and permutations evaluated by the ES algorithm increased at almost factorial rate from one level to the next. For example at level $k = 6$, the number of articulatory combinations considered for the search were over 2 million. The exhaustive search algorithm was implemented on the 1D and 2D models up to level 6 only, due to computational time constraints.

5.3 Results and evaluation

The results offer a comparison of the DFS and ES algorithms based on evaluation divergence, the critical articulator sets and the computational effort. The computed evaluation divergence was averaged across all phones to indicate which search procedure gave better overall fit to the articulatory data at each level, $k = 1..6$. As before, a range of critical threshold values was used ($0.1 \leq \theta_{\mathrm{C}} \leq 5$), to find the effect of the critical threshold on the search procedures' performance. For comparison of the consonant and vowel critical articulator sets for the 1D case, the convergence threshold was set to match the number of dimensions from IPA (ES at $\theta_C = 1.5$ for male and female; c.f. DFS at $\theta_C = 1.7$). A lower second threshold was chosen at double this number of critical articulators, $\theta_{\mathrm{C}} = 0.5$, which captured more detail (c.f. DFS at $\theta_C = 0.6/0.7$ for male/female). We refer to these two operating points as the IPA and lower thresholds respectively. Further decreasing the threshold showed only 10–12% improvement in the model convergence. Finally, we compare the two search techniques in terms of the computational effort.

Figure 10 shows the average evaluation divergence computed between model and phone pdfs as $0.1 \leq \theta_{\mathrm{C}} \leq 5$, for DFS and ES algorithms. For all values of convergence threshold, there was a negligible difference in their performance, both for 1D and 2D versions. There was a difference in the value of θ_{C} needed to yield any given number of critical dimensions, e.g., DFS had $\theta_{\mathrm{C}} = 1.7$ whereas ES had $\theta_{\mathrm{C}} = 1.5$ at IPA complexity. At the lower threshold $\theta_{\mathrm{C}} = 0.5$, the maximum *identification* divergence was reduced by 15% overall by ES over the proposed algorithm, but this reduction did not convert into better

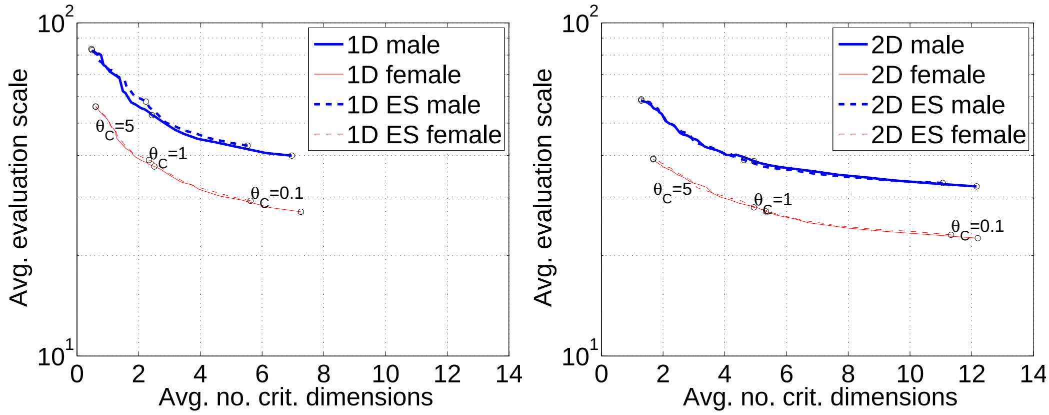


Fig. 10. Average 1D (left) and 2D (right) evaluation divergence between model pdfs and phone pdfs for proposed DFS (solid) and ES (dashed) algorithms with $\theta_C = \{0.1, 0.2, \ldots, 5.0\}$, for male (thick blue) and female (thin red) data.

performance. The change in evaluation divergence with the ES relative to DFS was small and even positive (+3%/2% for male/female). The differences were smaller for the 2D comparison. With only 0.5% decrease in evaluation divergence, averaged across both threshold levels and speakers, we found that ES provided a negligible benefit.

At the IPA threshold, there was good agreement: the two 1D critical articulator lists were identical for half the phones and, allowing permutations, the same for 60%/71% (male/female). In general, the order of critical articulator identification was found not to affect significantly the phone models' convergence. Treating any changes in the critical articulator lists from DFS to ES as errors, the overall identification accuracy was 74%. The largest proportion of discrepancies resulted from substitution of one or two articulators for strongly correlated ones. For instance, for female [s], the third critical articulator TBx was replaced by TDx, whereas the male list included the expected TTx with LIy and TTy. No other meaningful patterns were found in the comparison.

The ES algorithm was computationally expensive when compared with the proposed algorithm. It took 1.0×10^6 s (1D) and 5.8×10^3 s (2D) to run ES for one speaker upto level 6 (implemented in Matlab v7.5.0 on a machine with four 3.3 GHz processors with 32 GB RAM), whereas the DFS algorithm took less than 1 s in both 1D and 2D cases. Thus, the main difference between the two search strategies was the computational load. The proposed DFS algorithm was as effective as the ES method with an added advantage of faster execution times.

6 Implications and applications of the algorithm

6.1 Efficient articulatory modeling

In its current 1D and 2D forms, the critical articulator identification algorithm provides a compact representation of the articulatory configuration for each phone. The phone distributions were approximated here using the identity, statistics and correlations of critical articulators, with the grand statistics and correlations. To count the number of parameters, we use the number of articulators a=14 for 1D or a=7 for 2D, the model dimensionality K=1 for 1D or K=2 for 2D, and the number of phones n_Φ=44 in the inventory Φ. Thus, the means of phone pdfs require $an_\Phi K$ parameters, and the symmetrical (co)variances $an_\Phi \frac{K}{2}(K+1)$. The grand means, (co-)variances and inter-articulator correlations are needed to define the model pdfs, but then only the critical articulator statistics are needed for each phone. The grand statistics use $a\left(2K+\frac{K}{2}(aK-1)\right)$, and each phone uses $\gamma_C\left(2K+\frac{K}{2}(\gamma_C K-1)\right)$ where γ_C is the average number of critical articulators.

With complexity equivalent to the IPA descriptions, γ_C=1.8 for 1D and γ_C=1.4 for 2D, the reductions in the models' parameters were 80% and 77% respectively, averaged across the two subjects. The models became less compact as the number of critical dimensions increased, e.g., reductions in parameters of 61% and 28% were achieved at the lower threshold for 1D and 2D (with average 3.6 and 5.6 critical dimensions/phone). Thus, the representation from the proposed algorithm is more compact than a conventional, statistical description. Transformation of articulator coordinates may provide additional information compression, e.g., via principal or linear component decomposition (Maeda, 1990; Hoole et al., 2008; Badin et al., 2002). The authors have presented preliminary work in this direction (Jackson and Singampalli, 2008b). Yet here, the identified constraints of primary and secondary articulations are available for interpretation.

6.2 Potential impact in speech science and technology

Accurate statistical models of coarticulation are of prime importance for future advances speech synthesis and ASR. Such models aim to capture effects of target undershoot and overshoot, smooth and continuous articulator movement, and passive gestures of dependent articulators. The constraints encapsulated in the articulatory roles that our algorithm identifies can be used to prioritise speech gestures and to determine unconstrained degrees of freedom.[2] Allowing

[2] The latter relate to the uncontrolled manifold in gestural dynamic terms.

articulators to relax when redundant, for instance, could account for tongue dipping during bilabial VCV sequences.[3] A preliminary experiment by the authors relates our data-driven method to feature spreading: in synthesis of articulatory trajectories from phone labels (Singampalli and Jackson, 2007), modified redundant segments showed a small improvement over a baseline system that included inertial coarticulation effects. Extensions of the algorithm could derive gestural activation patterns of phone sequences from 2D or 3D articulatory data, rather than by phonetic rules. The quantitative description of articulatory contraints that we advocate provides structure of the phoneme-to-phone transformation in speech production, which can be used for improving context sensitivity.

In phonetics, critical articulations are typically specified by subjective feature-based or gestural descriptions; uncritical articulators have unspecified features or gestures, and different realizations are detailed using diacritics in 'narrow' phonetic transcriptions. Some variation can be explained using theories such as feature spreading (Henke, 1965; Moll and Daniloff, 1971; Daniloff and Hammarberg, 1973) and overlap of articulatory gestures (Browman and Goldstein, 1986; Saltzman and Munhall, 1989). The articulatory roles obtained using our algorithm can supplement these theories with objective, statistical evidence of significant coarticulation effects. It differentiates between critical articulator targets, consequent movements of linked parts of the anatomy, and redundant parts that are most susceptible to the biomechanical effects of coarticulation from targets of neighbouring phonemes. A brief phonetic analysis of identified articulatory constraints is presented in Jackson and Singampalli (2008a), that offers opportunities for modeling anticipatory and carry-forward coarticulation effects, based on MOCHA-TIMIT data. The algorithm can be used for linguistic studies of various languages, dialects and speakers, for instance in determining phonetic inventories.

In engineering, many ASR systems have attempted to incorporate articulatory constraints(King et al., 2007), inspired by distinctive features (Kirchhoff, 1999; Metze and Waibel, 2002; Frankel et al., 2004; Eide, 2001; Koreman et al., 1998), in the form of quantized gestural configurations (Deng and Sun, 1994; Erler and Freeman, 1996; Richardson et al., 2000), or within a hidden (pseudo-)articulatory layer via forward (Russell and Jackson, 2002; Richards and Bridle, 1999) or inverse mapping (Richmond, 2006; Frankel et al., 2000). The physiological constraints offered by human speech production have been incorporated into speech synthesis via articulatory codebooks, regression and neural-network approaches for forward mapping from articulatory to acoustic domains, as in (Schroeter and Sondhi, 1994). It is widely accepted that appropriate use of articulatory information provides constraints that can improve

[3] The tongue acts like a lazy ballerina, drooping to save effort while the spotlight is on another dancer and recovering her pose when it shines on her again.

the performance of speech technologies; the remaining open problem concerns how to do so. The present research contributes in identifying essential gestural commands in the planning and articulation of real speech utterances.

7 Conclusion

In this paper, we have proposed an algorithm for identifying critical, dependent and redundant roles played by articulators during speech production. The algorithm finds critical articulators using the identification KL divergence between the phone and grand distributions, and updates pdfs of dependent articulators based on their overall correlation with the critical articulators. The 1D and 2D version of the method were applied to EMA data from MOCHA-TIMIT. Results were analysed and compared to IPA descriptions of the speech sounds. The accuracy of fit to the measured phone distributions was evaluated by computing the evaluation divergence (with full phone covariance), across a range of thresholds. As expected, the 1D models were outperformed at all threshold values by the 2D ones that modeled the covariance between x and y movements. These models also fitted the measured distributions better than ones derived from IPA descriptions of phoneme articulation, despite considerable benefit from incorporating the biomechanical dependencies between the articulators.

Phonetic analysis showed that the algorithm output compared well to IPA descriptions for consonants, while fricatives claimed additional critical articulators. It distinguished between full and central or reduced vowels, whose configurations were more susceptible to coarticulation and had no critical articulator. Some insertions and substitutions occurred where there was strong correlation between articulators, and various speaker differences were seen. In evaluation of the proposed algorithm against an exhaustive search, where all critical articulator combinations were tested according to a minimax criterion, we found that the proposed method performed as well as the exhaustive search for much less computation.

The model of phone distributions obtained using the proposed algorithm, through recognition of articulatory roles, is shown to be more compact and more informative than a conventional statistical description. Applications that exploit models of coarticulation and trajectory generation for audio and visual speech synthesis and recognition abound, offering plenty of scope for development of data-driven approaches to mapping the relationship between phonemes and their realization as phones. In the field of phonetics, knowledge of real articulatory roles has potential for explaining coarticulation effects, and studying phonetic inventories for different languages, speakers and styles. One attempt at generating synthetic trajectories from phone labels was mentioned that gave encouraging results. Further work is needed to investigate dynamic behaviour of articulatory constraints. Opportunities exist to extend critical articulator analysis to other types of speech data, and to explore knowledge of articulatory roles in the synthesis of speech, whether explicitly, e.g., for visual/articulatory speech synthesis, or implicitly, e.g., in a join cost or smoothing function for concatenative synthesis. Our interest focuses on ways of exploiting new knowledge of articulatory constraints as conditional dependencies in probabilistic speech models for ASR.

References

Anderson, T., 1984. An introduction to multivariate statistical analysis, 2nd Edition. Wiley, New York.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Savariaux, C., 2002. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. J. Phon. 30 (3), 533–553.

Badin, P., Serrurier, A., 2006. Three-dimensional linear modeling of tongue: Articulatory data and models. In: Laboissière, R. (Ed.), Proc. Int. Sem. Spch. Prod. (ISSP'06). Ubatuba, Brazil, pp. 395–402.

Bakis, R., 1991. Coarticulation modeling with continuous-state HMMs. Proc. IEEE Workshop Automatic Speech Recognition, Harriman, New York, 20–21.

Blackburn, S., Young, S., March 2000. A self-learning predictive model of articulator movements during speech production. J. Acoust. Soc. Am. 103 (3), 1659–70.

Bladon, R. A. W., Al-Bamerni, A., 1976. Coarticulation resistance in English /l/. J. Phon. 4, 135–50.

Browman, C. P., Goldstein, L., 1986. Towards an articulatory phonology. Phonology 3, 219–52.

Chomsky, N., Halle, M., 1968. The sound pattern of English. Harper & Row, New York.

Cohen, M. M., Massaro, D. W., 1993. Modeling coarticulation in synthetic visual speech. In: Models and Techniques in Computer Animation. Springer-Verlag, pp. 139–156.

Coker, C. H., 1976. A model of articulatory dynamics and control. Proc. IEEE 64 (4), 452–460.

Coppin, B., 2004. Artificial Intelligence Illuminated, 1st Edition. Jones & Bartlett, ISBN 0763732303.

Dang, J., Honda, K., 2004. Construction and control of a physiological articulatory model. J. Acoust. Soc. Am. 115 (2), 853–870.

Dang, J., Honda, M., Honda, K., 2004. Investigation of coarticulation in continuous speech of Japanese. Acoust. Sci. & Tech. 25 (5), 318 – 329.

Dang, J., Wei, J., Suzuki, T., Perrier, P., 2005. Investigation and modelling of coarticulation during speech. Proc. Interspeech, *Lisbon*, 1025–1028.

Daniloff, R., Hammarberg, R., 1973. On defining coarticulation. J. Phon. 1, 239–248.

Deng, L., Sun, D. X., 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J. Acoust. Soc. Am. 95 (5), 2702–2719.

Eide, E., 2001. Distinctive features for use in an automatic speech recognition system. Proc. Eurospeech '01, *Aalborg, Denmark*, 1613–1313.

Erler, K., Freeman, G. H., 1996. An HMM-based speech recognizer using overlapping articulatory features. J. Acoust. Soc. Am. 100 (4), 2500–2513.

Fant, G., 1969. Distinctive features and phonetic dimensions. Applications of Linguistics, Cambridge, UK.

Frankel, J., 2003. Linear dynamic models for automatic speech recognition. Ph.D. thesis, CSTR, Univ. of Edinburgh.

Frankel, J., King, S., 2001. ASR-Articulatory Speech Recognition. Proc. Eurospeech '01, *Aalborg, Denmark*, 599–602.

Frankel, J., Richmond, K., King, S., Taylor, P., 2000. An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. Proc. Int. Conf. on Spoken Lang. Proc., *Beijing* 4, 254–257.

Frankel, J., Wester, M., King, S., 2004. Articulatory feature recognition using dynamic Bayesian networks. Proc. Int. Conf. on Spoken Lang. Proc., *Jeju, Korea*, 1477–1480.

Henke, W. L., 1965. Dynamic articulatory model of speech production using computer simulation. Ph.D. thesis, MIT, Cambridge, MA.

Hershey, J. R., Olsen, P. A., 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. Proc. IEEE-ICASSP 4, 317–320.

Hoole, P., Wismueller, A., Leinsinger, G., Kroos, C., Geumann, A., Inoue, M., 2008. Analysis of the tongue configuration in multi-speaker, multi-volume MRI data. In: Proc. Int. Sem. Spch. Prod. (ISSP'00). Kloster Seeon, Germany, pp. 157–160.

Jackson, P. J. B., Singampalli, V., Shiga, Y., Russell, M. J., 2004. Dansa project: Statistical models to relate speech gestures to meaning. CVSSP, Univ. of Surrey, Guildford, UK, EPSRC GR/S85511/01 [http://www.ee.surrey.ac.uk/Personal/P.Jackson/Dansa/].

Jackson, P. J. B., Singampalli, V. D., Dec. 2008a. Coarticulatory constraints determined by automatic identification from articulograph data. In: Proc. Int. Sem. on Spch. Prod. (ISSP'08). Strasbourg, France, pp. 377–380.

Jackson, P. J. B., Singampalli, V. D., 2008b. Statistical identification of critical, dependent and redundant articulators. J. Acoust. Soc. Am. 123 (5, Pt. 2), 3321, Presented at Acoustics'08, Paris.

Johnson, R. A., Wichern, D. W., 1998. Applied multivariate statistical analysis, 4th Edition. Prentice Hall, New Jersey.

Keating, P. A., 1988. The window model of coarticulation: articulatory evidence. UCLA Working papers in Phonetics 69, 3–29.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., February 2007. Speech production knowledge in automatic speech recognition. J. Acoust. Soc. Am. 121 (2), 723–742.

Kirchhoff, K., 1999. Robust speech recognition using articulatory information. Ph.D. thesis, Univ. of Bielefeld.

Koreman, J., Andreeva, B., Barry, W. J., 1998. Do phonetic features help to improve consonant identification in ASR? Proc. Int. Conf. on Spoken Lang. Proc., *Sidney, Australia* 3, 1035–1038.

Kullback, S., 1968. Information theory and statistics, 1st Edition. Dover Pub., New York.

Liberman, A. M., 1970. The grammars of speech and language. Cog. Psych. 1, 301–23.

Lindblom, B., 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773–81.

Löfqvist, A., 1990. Speech as audible gestures. In: Hardcastle, W., Marchal, A. (Eds.), Speech production and Speech Modeling. Kluwer Academic Publishers, pp. 289–322.

MacNeilage, P. F., 1970. Motor control of serial ordering of speech. Psychol. Rev. 77, 182–196.

Maeda, S., 1990. Speech Production and Modelling. Kluwer, Ch. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model, pp. 131–149.

Meister, I. C., Wilson, S. M., Deblieck, C., Wu, Allan D. amd Iacoboni, M., 2007. The essential role of premotor cortex in speech perception. Current Biology 17 (19), 1692–1696.

Mermelstein, P., 1973. Articulatory model for the study of speech production. J. Acoust. Soc. Am. 53 (4), 1072–1082.

Metze, F., Waibel, A., 2002. A flexible stream architecture for ASR using articulatory features. Proc. ICSLP, *Denver, CO*, 2133–2136.

Moll, K., Daniloff, R., 1971. Investigation of the timing of velar movements during speech. J. Acoust. Soc. Am. 50 (2), 678–84.

Öhman, S. E. G., 1966. Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Am. 39 (1), 151–68.

Öhman, S. E. G., 1967. Numerical model of coarticulation. J. Acoust. Soc. Am. 41 (2), 310–20.

Ostry, D. J., Gribble, P. L., Gracco, V. L., 1996. Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? The Journal of Neuroscience 16 (4), 1570–1579.

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., Levy, S., 1992. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. J. Acoust. Soc. Am. 92 (2), 688–700.

Parthasarathy, V., Prince, J. L., Stone, M., Murano, E. Z., NessAiver, M., 2007. Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing. J. Acoust. Soc. Am. 121 (1), 491–504.

Qin, C., Carreira-Perpiñán, M., Richmond, K., Wrench, A., Renals, S., Sep. 2008. Predicting tongue shapes from a few landmark locations. In: Proc. Interspeech. Brisbane, Australia, pp. 2306–2309.

Recasens, D., Pallarés, D., 1999. A study of /ɾ/ and /r/ in the light of the 'DAC' coarticulation model. J. Phon., 143 – 170.

Recasens, D., Pallarés, D. M., Fontdevilla, J., 1997. A model of lingual coarticulation based on articulatory constraints. J. Acoust. Soc. Am. 102 (1), 544–561.

Richards, H. B., Bridle, J. S., 1999. The HDM: A segmental hidden dynamical model of coarticulation. Int. Conf. on Acoustics, Speech, and Signal Processing, *Phoenix, Arizona, USA* 1, 357–360.

Richardson, M., Blimes, J., Diorio, C., 2000. Hidden-articulatory Markov models for speech recognition. Proc. Int. Conf. on Spoken Lang. Proc., *Beijing* 3, 131–134.

Richmond, K., 2006. A trajectory mixture density network for the acoustic-articulatory inversion mapping. Proc. Interspeech, *Pittsburgh, PA*.

Richmond, K., 2007. A multitask learning perspective on acoustic-articulatory inversion. In: Proc. Interspeech. Antwerp, Belgium, pp. 2465–2468.

Russell, M. J., Jackson, P. J. B., 2002. Models of speech dynamics in a segmental-HMM recogniser using intermediate linear representations. Proc. ICSLP, *Denver, CO*, 1253–1256.

Saltzman, E. L., Munhall, K., 1989. A dynamic approach to gestural patterning in speech production. Ecology Psychology 1 (4), 333–82.

Schroeter, J., Sondhi, M. M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trans. SAP 2 (1), 133–150.

Shadle, C. H., 1985. The acoustics of fricative consonants. Ph.D. thesis, MIT, Cambridge, MA.

Singampalli, V. D., Jackson, P. J. B., 2007. Statistical identification of critical, dependent and redundant articulators. Proc. Interspeech, *Antwerp*, 70–73.

Soquet, A., Saerens, M., Lecuit, V., 1999. Complementary cues for speech recognition. Proc. ICPhS, *San Francisco, CA*, 1645–1648.

Tokuda, K., Zen, H., Kitamura, T., 2007. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic vector feature sequences. Comp. Speech & Lang. 21 (1), 153–73.

Westbury, J. R., Turner, G., Dembowski, J., 1994. X-ray microbeam speech production database user's handbook. Univ. of Wisconsin, Madison, WI.

Wilson, S. M., Saygun, A. P., Sereno, M. I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nature Neuroscience 7, 701–702.

Wrench, A. A., 2001. A new resource for production modelling in speech technology. Proc. Inst. of Acoust., *Stratford-upon-Avon, UK* 23 (3), 207–217.