Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 45 (2017) 253–277

www.elsevier.com/locate/csl

Öhman returns: New horizons in the collection and analysis of imaging data in speech production research☆

Philip Hoole*, Marianne Pouplier

Institute of Phonetics and Speech Processing, LMU, Schellingstr. 3, 80799 Munich, Germany

Received 31 August 2016; Accepted 3 March 2017

Abstract

There have been enormous technical advances in the use of imaging techniques in speech production research in terms of resolution and frame rates. However, a major bottleneck lies in the lack of appropriate data reduction and quantification methods which allow for a parsimonious representation of the high-dimensional image data. Particularly the rapid increase in frame rates seen in data acquisition makes traditional, error-prone methods of contour tracking unwieldy due to the high amount of manual intervention required. We discuss recent developments in methods that obviate contour tracking but instead process the entire image. Specifically, we focus on one such approach by demonstrating the application of Principal Component Analysis to ultrasound images. This method not only exploits the information present in the entire image, but it also straightforwardly allows for the representation of the temporal evolution of an utterance by reducing a series of images to a time-varying single-value representation. In an illustrative example, inspired by the seminal work of Öhman (1966), we show how characteristic patterns of coarticulation between vowels and consonants can be captured in this way.

Keywords: Articulatory data; Imaging techniques; Time series analysis; Ultrasound; Principal component analysis; Coarticulation

1. Introduction

Over the past 10 years a rapidly growing number of speech science laboratories have used imaging technology to study speech articulation. Medical ultrasound (US) systems are becoming more affordable and portable and are increasingly used to study tongue motion during speech. Also real-time MRI is on the rise in phonetic research, with a number of publications showcasing ever improving spatio-temporal resolution (e.g., Fu et al., 2015; Narayanan et al., 2004; Niebergall et al., 2012; Uecker et al., 2010). Particularly ultrasound has enjoyed recent popularity since it is cheap compared to MRI, portable, non-invasive, and safe; moreover, it can be used with populations difficult to study otherwise such as children and patients. In that sense, ultrasound affords the opportunity for a democratization of speech production research, which has so far largely been restricted to comparatively few technically high-powered phonetics labs in the world. In addition, the cost- and time-efficiency of ultrasound recordings in principle

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.
E-mail address: hoole@phonetik.uni-muenchen.de (P. Hoole), pouplier@phonetik.uni-muenchen.de (M. Pouplier).

http://dx.doi.org/10.1016/j.csl.2017.03.002
0885-2308/

allows for articulatory studies with representative sample sizes,[1] yet data quantification is still a major bottleneck in this respect. Efficient data quantification methods and adequate statistical evaluation of imaging data taking into account the complex correlation structures typical for phonetic experiments are only in their beginnings. Particular problems associated with imaging data in general lie in the sheer size of the image data (making increasing frame rates a mixed blessing in this respect) as well as in the difficult task of extracting relevant information on tongue movement from noisy images. The complexities of data analysis have restricted many studies to making largely qualitative observations on a few speakers or employing overly simplifying data reduction methods. For this reason, also following a trend in the speech sciences generally, more and more researchers have devoted great efforts to the development of advanced data processing methods and adequate statistical tests taking complex study designs into account (Barbosa et al., 2012,2008; Cederbaum et al., 2016; Gubian et al., 2015; Lancia et al., 2015; Lancia and Tiede, 2012; Wieling et al., 2014). The current paper, too, seeks to contribute to these developments by illustrating how data analysis techniques based on the entire image offer a promising avenue for exploiting 'big data' in speech production research. We apply our principal-component-based method to ultrasound data here, but it can generally be applied to any kind of image data.

1.1. Ultrasound in speech production research

Speech ultrasound data are acquired using medical ultrasound systems, usually in B-Mode (for an overview, see Stone, 2005). The high-frequency sound waves emitted by the piezoelectric crystals located in the ultrasound probe are reflected, among others, by the tissue–air boundary at the tongue surface. Intensity and time lag of the returning echo are used to generate a time series of grey- or color-scaled images depicting, in this case, the tongue. For data quantification a contour is usually fitted semi-algorithmically to the outline of the tongue surface (Fig. 1 left). Contours are typically extracted based on an equidistant-angle polar grid superimposed on the image (e.g., Wrench, 2008); alternatively active contour or snake algorithms are used in a Cartesian system (Iskarous, 2005a; Li et al., 2005). Due to noise, the traceable portion of the tongue may vary from frame to frame; problematic data are usually discarded. Each ultrasound image thus renders a tongue shape curve (Fig. 1 right).

While it is true that the automatic extraction of the contours of the articulators is becoming increasingly sophisticated (Berry and Fasel, 2011; Fasel and Berry, 2010; Silva and Teixeira, 2015) it is typically the case that analysis of imaging data can involve a lot of manual intervention. Semi-automatic tracking or segmentation algorithms developed for imaging data generally (Bresch and Narayanan, 2009; Iskarous, 2005a; Li et al., 2005; Proctor et al., 2010; Wrench, 2008) show an error margin large enough that manual frame-by-frame correction is necessary. Given that only around 5–10 speakers are typically recorded in imaging studies, tracking errors can severely distort the results. The problem of time-consuming manual correction has become all the more acute with increasing frame-rates: Even

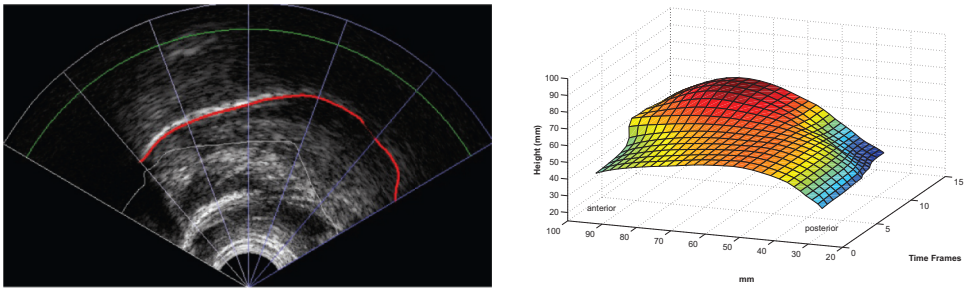

Fig. 1. Left: Single ultrasound frame of the tongue during the vowel /a/ with superimposed polar grid and tongue contour (Articulate Assistant Advanced; Wrench, 2008). Tongue tip is to the left. Right: Time series of tongue contours from the same speaker, utterance 'padi'. 15 equidstant frames were extracted during the medial VCV sequence from the beginning of /a/ to the end of /i/.

[1] At present articulatory studies typically include no more than 5–10 speakers, rarely more, often less, irrespective of the method employed. This is due to the high costs and labor-intensive nature of the available experimental methods. Inter-speaker variability in speech production research has thus been a notorious problem for generalizability and replicability of study results.

though ultrasound can now better capture most articulatory processes as a result of the doubling or even trebling of acquisition frame rates, it is still often the case that data analysis is based on one or two selected frames, discarding most of the acquired data (among many others, Lin et al., 2014; Mielke, 2015). Another notorious problem of imaging data is that even single time point analyses pose a challenge to data quantification: if tracking software is used to outline the contour of the tongue as has widely been done particularly in X-ray as well as ultrasound research (Davidson, 2005; Iskarous, 2005b; Öhman, 1966; Pouplier, 2008), every analysis point represents a curve. This requires further data reduction techniques even for approaches based on a few selected frames ('magic moments' in the sense of Mücke et al., 2014). Some studies chose single points of maximal excursion (Pouplier, 2008; Proctor, 2011; Yip, 2013), while others have discussed various shape-capturing techniques (Davidson, 2006; Dawson et al., 2016; Ménard et al., 2012).

In this paper we present an approach that seeks to circumvent the problem of curve extraction and parametrization altogether by applying data reduction in the form of Principal Component Analysis (PCA) to the entire image (for general background to the use of PCA in articulatory studies, see e.g. Badin et al., 2002). Tongue imaging data from ultrasound are used for demonstrating how the temporal structure of spoken utterances, specifically coarticulatory patterns, can be analyzed this way.

1.2. Processing speech imaging data: PCA and other approaches

PCA is a non-parametric dimension reduction technique designed to extract patterns from high-dimensional multivariate data by finding orthogonal axes of variation in terms of eigenvectors and their associated eigenvalues. PCA performs a linear transformation of the variables into a lower dimensional space while retaining a maximal amount of information about the variables. The data are projected into a new (lower dimensional) coordinate space with each new coordinate accounting for, in decreasing order, the next greatest amount of variance. The eigenvectors (principal components, PCs) of the covariance matrix correspond to the directions of greatest variance. The eigenvalues give the amount of variance which is accounted for by the corresponding eigenvector. The eigenvector with the largest eigenvalue is the direction along which the data set has the maximum variance. Principal components are linear combinations of the original variables weighted by their contribution to explaining the variance in a particular dimension. Thereby it is important to keep in mind that PCA partitions the variability patterns in the data among orthogonal axes of variation independently of any independent variables in the dataset (in this respect PCA is crucially different from Linear Discriminant Analysis, for example, which derives linear combinations of the original variables to maximize classification accuracy with respect to independent variables in the dataset). Straightforward interpretability of the PCs is not guaranteed, a point we shall return to in Section 4.

As we discuss in this paper, PCA analysis of the several thousand pixels of which each ultrasound image is composed provides one potentially attractive alternative to contour tracking. PCA approaches specifically are motivated by a rich phonetic literature consistently showing that the tongue has about four or five degrees of freedom, i.e. the range of tongue shapes found in speech can be expressed as a weighted sum of 4–5 underlying functional 'building blocks' (Beautemps et al., 2001; Harshman et al., 1977; Harshman and Lundy, 1984; Hoole, 1999; Maeda, 1990). The advantage of processing the whole tongue image is that such an approach can take advantage of all of the rich information contained in the image, ranging from the floor of the mouth up to the surface of the tongue – information that cannot be obtained from techniques such as x-ray. In fact, the tissue–air interface at the tongue surface is often quite poorly defined (making the contour tracking described above a tough problem), while the orientation of the tongue muscle fibres may still clearly be visible. It is precisely because the tongue – often referred to as a muscular hydrostat, like an elephant's trunk – consists largely of incompressible muscle tissue that the surface of the tongue must reflect in a coherent way changes over the whole volume of the organ, for instance in terms of the orientation and compression of muscle fibres.

The application of PCA to image data has been around for some time, including the groundbreaking work of Turk and Pentland (1991) on facial recognition, and the analysis of the tongue by Hueber et al. (2007). However, in both of these studies the emphasis has been on PCA as a coding and classification tool (in the latter paper in the context of a silent speech interface). Thus the thrust of our own work is in a sense closer to that of Lammert et al. (2010) and Proctor et al. (2013,, 2011) on MRI data in which they explore ways of automatically extracting time series measures that correspond to features like constriction degree for key articulators. For example, they showed how simply calculating average pixel intensity in a suitable region of interest can capture the approximation of the moving articulator

to the immoveable vocal tract wall since (in MRI) forming a constriction involves more high-intensity tongue tissue, and less low-intensity air in the region of interest. Optical flow is another method that extracts information from the entire image and that has recently been used to analyze speech data (Barbosa et al., 2012,, 2008), for application to ultrasound data see Hall et al. (2015) and Moisik et al. (2014). This technique evaluates the direction and magnitude of frame-to-frame shifts in pixel intensity patterns (essentially pixel velocities). The output of optical flow analysis is, however, high-dimensional and requires further data reduction techniques such as PCA or reducing the size of the analysis space by defining regions of interest within the image. For other approaches to entire image processing, see Jacobs et al. (2008), Lancia et al. (2015) and McMillan and Corley (2010).

PCA can be seen as an extension of these types of pixel intensity-based approaches since it defines a weighting scheme that is applied to the pixels in an image (or region of interest) that will capture position along a characteristic axis of variation. Carignan (2014), building on the work of Hueber et al. (2007), has recently presented a series of Matlab algorithms to perform PCA analyses on entire ultrasound images. Using this technique, Carignan et al. (2016) extracted PC scores from series of ultrasound images. They retained 20 PCs to describe the ultrasound images. While the large number of PCs may also be due to the varied speech material used in that study, this also highlights the different nature of ultrasound data compared to e.g. the x-ray data the PARAFAC model (Harshman and Lundy, 1984) was based on: The noisy nature of the ultrasound image together with the fact that no other anatomical structures apart from the tissue−air interface of the tongue surface (together with reflections of the internal structure of the tongue) are visible, drastically increase the number of PCs needed to adequately capture the image. This of course is a challenge to data interpretation. Carignan et al. achieve interpretability of their 20 PC scores by relating them via linear regression to acoustic formant distance values on a sample-by-sample basis. The result is a time series of a scalar variable that can be interpreted as tongue height (cf. Carignan et al., 2015, for an analysis of MRI data in a similar vein). In the present paper, we similarly discuss the potential of PCA for extracting time series information from ultrasound images.

In sum, given the noisiness of ultrasound images, and the fact that tongue contours are often not well defined (particularly when the contour is oriented parallel to the scan lines of the transducer) it seems reasonable to make use of all coherent sources of information in an image. Methods that process the entire ultrasound image are a particularly promising avenue for the use of imaging data in phonetics since they avoid very labor-intensive contour tracking. The methods will thus scale better to larger studies and to evaluation of the full temporal sequence of interest. As a consequence, data loss and error due to tracking failure will be reduced. While data using the entire images poses 'big data' problems − potentially 1 billions of data points per subject − PCA, as we will demonstrate below, provides a parsimonious approach to data reduction while preserving the dynamic information on tongue shape changes. We will now give a brief description of the research context in which the current data were acquired.

1.3. The current study: Öhman returns

The data we will analyze in this paper were recorded in order to investigate coarticulation resistance in German coronals, in particular /l/. Laterals of the world's languages are traditionally classified along a 'light-dark' continuum in relation to the degree of tongue dorsum involvement (Ladefoged and Maddieson, 1996; Recasens, 2012). The German lateral is considered to be a clear [l] and thus to have a dorsal component that is small or absent. The degree of dorsal control has generally been associated with coarticulation resistance (Recasens and Espinosa, 2009) such that consonants with a high degree of dorsal control have great coarticulation resistance and minimal variation with context (Bladon and Al-Bamerni, 1976). On the basis of ultrasound data from Spanish and Russian, Proctor (2011) concluded that laterals in many languages, whether clear or dark, exert dorsal control, making them more coarticulation resistant than stops. This stands in contrast to articulography data on German by Geumann et al. (1999): For two out of their four speakers there was no difference in vowel-conditioned variability between /d, l/ (as measured by positional variation of the sensor attached to the tongue dorsum), for one speaker /l/ showed a higher degree of constraint compared to /d/ as predicted by Proctor (2011), while the fourth speaker showed the opposite pattern (thereby nicely illustrating the problem of generalizing results across speakers based on articulatory data − see Footnote 1). It is problematic though to pit the Geumann et al. and the Proctor results against each other, since it is not entirely clear in which area of the tongue synergistic control of the tongue back for lateral production would restrict vowel coarticulation the most (cf. especially Fig. 1 in Recasens, 2012, and Chapter 6 'Laterals' in Ladefoged and Maddieson, 1996). Whether these synergistic effects are found may depend on how the articulography sensors are positioned

on the tongue – this type of artifact may, in turn, contribute to some of the inter-subject variability observed by Geumann and colleagues. Holistic evaluation of tongue shape changes, as is possible with ultrasound, seems to be ideally suited to advance our understanding of the relative degree of coarticulatory variation allowed for by laterals. Our dataset contains German coronals in varying vowel contexts allowing us to look at differential degrees of vowel-conditioned variation between the consonants as well as the degree of vowel-to-vowel coarticulation each consonant allows for. We use this data set to illustrate the potential of PCA for the analysis of ultrasound data. Since our focus in the present study is primarily methodological, we will mainly demonstrate how our method can be used to uncover general patterns of vowel and consonant coarticulation instead of strictly focusing on a single research question. Being able to confirm independently known coarticulatory patterns in our data is an important part of methodological validation in this context.

2. Methods

2.1. Ultrasound image acquisition and pre-processing

Ultrasound data was collected by means of an Ultrasonix SonixTouch scanner using a C9-5/10 microconvex (radius 10 mm) transducer with a transmitter frequency of 5 MHz and a maximum field of view of 148 deg. The probe was held in place using the Articulate Instruments helmet-style probe-holder (Scobbie et al., 2008). For each speaker, the probe placement and ultrasound settings were adjusted such that the shadow of the jaw and of the hyoid bone, respectively, constituted the right and left edges of the image (Stone, 2005). During each trial (lasting a few seconds) the imaging data was stored locally on the scanner and then transferred digitally using a TCPIP connection to a host computer running the ArticulateAssistantAdvanced software package (Articulate Instruments) by Wrench (2008). A frame sync signal generated by the scanner was used to align the image data with simultaneously recorded audio data; a Sennheiser MKH40 microphone was used.

The image generated by a microconvex probe consists of a set of scanlines fanning out from the origin, with a given number of pixels located equidistantly along each scan line (see left panel of Fig. 2). For all subjects except S1 we used a setting of 128 scanlines for the full field of view (for S1 the scanner was inadvertently set to use a higher number of scanlines). By excluding scanlines for which the tongue was shadowed by the mandible or hyoid, typically 80–100 scanlines were actually acquired per speaker, allowing a corresponding increase in frame-rate. It may seem unusual to have frame rates differing by subject, yet this is due to the data acquisition rate in ultrasound

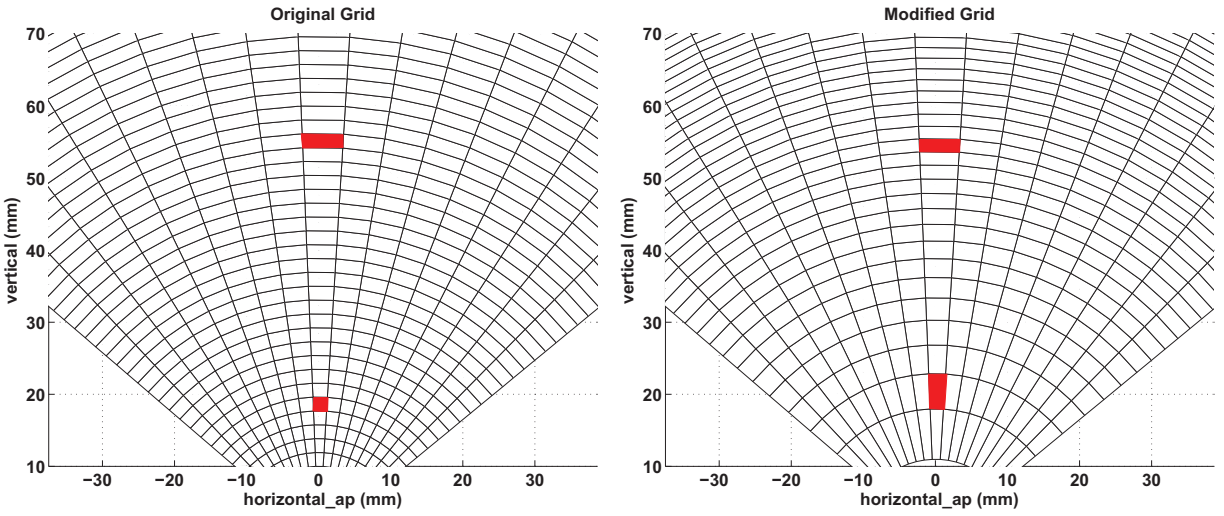

Fig. 2. Illustration of the grid defining pixel locations in the original image (left), and in the modified arrangement used as input to the analysis (right). Pixels are located at all points of intersection in the grid. Note that in the original grid the area of the quadrilaterals defined by sets of four adjacent pixels becomes larger as distance from the probe increases. In the modified grid the distribution of pixels along each scanline is adjusted to give equal-area quadrilaterals. Representative quadrilaterals at two different distances from the probe are highlighted by red shading in each panel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

depending on the scan settings (depth, number of scanlines, size of the sector scanned). Scan settings are usually custom fit to optimally capture a given speaker's vocal tract, i.e. they are adjusted to each speaker's vocal tract size. If, as in many ultrasound set-ups, only video data are exported from the ultrasound machine, the actual scan rate is mapped by some internal mechanism to the fixed video-frame rate (29.97 fps for NTSC format, 25 fps for PAL). Constant frame-rates across subjects are in these types of system an artifact of the conversion to video format. In our setup, we export the pre-scan-converted data via direct digital transfer from the scanner and thus the image is not reconstructed as a video signal (this procedure is also much more efficient in terms of storage: a video reconstruction of the ultrasound image might typically have of the order of 480 × 640 pixels, but of course contains no more information than the native scanner format in which, as will be detailed below, each image is of the order of 400 × 90 pixels). See Wrench and Scobbie (2006) for further discussion of the artifacts and disadvantages associated with converting ultrasound images to typical video formats and frame-rates. The analysis method we present below (see Section 2.3) is, however, applicable to video images as well.

For all subjects except S7, the pixel resolution along each scanline amounted to 0.192 mm. With a depth setting of 80 mm (S2, S3, S5) this gave 416 pixels per scanline. The anatomy of subjects S1, S4 and S6 allowed a depth setting of 70 mm resulting in 364 pixels per scanline and a corresponding increase in frame-rate. For S7 a slightly higher sampling rate for the radio-frequency data was used, giving slightly higher spatial resolution along the scanlines (520 pixels at a depth setting of 80 mm).

These basic settings are summarized in Table 1.

Prior to entering the raw image data of m pixels × n scanlines into the PCA the following adjustments were made: Due to the microconvex design of the probe, the distance between corresponding pixels in adjacent scanlines is much less at locations close to the probe compared to further away. For a subject with typical settings like S3 this means that the area of the quadrilateral enclosed by pixels m and $m+1$ in scanlines n and $n+1$ increases from approximately 0.05 mm^2 at locations closest to the probe to 0.35 mm^2 at a depth of 80 mm. See the left panel of Fig. 2 for illustration (note that to give a clearer view of the grid in the figure the scanlines have been subsampled by a factor of five, and the number of pixels along each scanline has been subsampled by a factor of 10). While one of the main motivations for a procedure such as PCA is that it makes it possible to take into account coherent patterns of tongue deformation over the whole body of the tongue (rather than just focusing on the tongue surface), simply using the raw set of pixel data would in effect very strongly weight the contributions of the parts of the tongue closest to the probe because of the much larger number of pixels there. Accordingly, the distribution of pixels along each scanline was changed such that the area of each quadrilateral as defined above stayed the same for all distances from the probe (illustrated in the right panel of Fig. 2). For this reference area we chose a value of 0.2 mm^2. For subjects 2–6 this corresponded to the area at a depth of 40 mm. This means that for positions along each scanline below 40 mm the interpixel distance was increased, and decreased for positions above 40 mm. This reference value was chosen such that the image resolution after re-mapping would be close to its original resolution in regions where the tongue surface is typically located. Re-mapping with this reference value was applied to all subjects (S1 was recorded with a smaller distance between scanlines, and S7 with a smaller distance between pixels) so that the data used in the analysis had nominally the same areal resolution for all subjects (and roughly the same number of pixels per frame for all subjects; see below and Table 2).

We compared the PCA results both with and without this re-mapping. The results were actually very similar, the first component being virtually indistinguishable on visual inspection. Typically, with the re-mapped data the

Table 1
Ultrasound acquisition parameters by speaker.

	Frames per s.	Radial resolution (deg.)	Scanlines used	Pixel resolution (mm)	Depth (mm)
S1	45.8	0.58	192	0.192	70
S2	91.8	1.16	86	0.192	80
S3	91.8	1.16	86	0.192	80
S4	102.3	1.16	86	0.192	70
S5	91.8	1.16	86	0.192	80
S6	97.8	1.16	90	0.192	70
S7	76.8	1.16	102	0.154	80

Table 2
Details of principal component analysis.

	Pixels used in PCA	# training frames	PC1 %var	PC2 %var	PC3 %var
S1	17,500	126	21.2	7.4	4.5
S2	17,500	258	27.1	5.1	3.6
S3	22,500	167	18.3	6.2	4.7
S4	17,500	239	13.6	7.3	4.7
S5	21,000	177	26.7	7.6	5.6
S6	19,000	229	17.0	4.5	3.4
S7	22,500	191	24.4	9.9	6.5

variance explained by the first component was higher by a few percent (but this may be just because re-mapping generally resulted in slightly fewer pixels being used in the analysis). The observation that the uncorrected data, in which greater weight is given to regions furthest away from the surface, did not cause greatly different results from the data with normalized inter-pixel distance renders in fact great support to our view that tongue surface changes must be reflected in coherent co-variation in the lower regions of the image, meaning the internal structures of the tongue.

2.2. Speakers and corpus

Seven native speakers of German (3 males, 4 females) volunteered to participate in the experiment. Speakers were pre-screened for image quality, as is generally customary for ultrasound research (Stone, 2005).[2] All but two of the speakers were trained phoneticians (S4, S5), but all participants were naive as to the purposes of the experiment.

The corpus spoken by the subjects consisted of nonsense words of the form /pVCV/ embedded in the phonetically neutral carrier phrase "Gebe ____ besser ab" (*Better give ... away*). The carrier phrase was chosen to have minimal coarticulatory impact on the target word. C was one of the 6 consonants /b, d, l, n, z, ʃ/ (for S1 /ʃ/ was not recorded). The labial condition served as basis for setting up the PCA model (see Section 2.3); /ʃ/ was included as a 'reference' consonant with a high degree of overall tongue shape control (Toda, 2009). The consonants were combined with the following 7 vowel contexts: /a_a/, /a_ə/, /a_i/, /e_a/, /e_ə/, /i_a/, /i_i/. For S1 only the vowel contexts /a_a/, /a_i/, /i_a/, /i_i/ were recorded. Stress was on the first syllable in accordance with the default German stress pattern. German /i/ and /e/ in stressed position are close to the corresponding cardinal vowels. /a/ in stressed position is typically a central vowel between Cardinal Vowels 4 and 5. In the unstressed V2 position in our material it was generally closer to a low schwa [ɐ].

The material was recorded in randomized order (together with a small amount of additional material that will be briefly touched on below). The number of repetitions recorded was seven for Subjects S3, S4, S6 and S7, six for S2 and S5, and eight for S1 (for whom the corpus was slightly smaller).

2.3. Ultrasound analysis

Following synchronization of the audio signal with the ultrasound data the MAUS system for automatic segmentation and labeling (Kisler et al., 2016; Schiel, 1999) was used to determine the temporal intervals corresponding to V1, C, and V2 in the audio signal. Segment boundaries were cross-checked using sonagram and waveform displays, and were corrected manually if necessary. The onset of V1 was located at the onset of voicing following the aspirated /p/ of the carrier word. The offset of V2 was located at closure onset of the following /b/ in the carrier phrase. The boundaries between V1 and C, and between C and V2 were located at the start and end of C closure, respectively, as visible in the sonagram.

As mentioned in the Introduction, a challenge of PCA analysis is in creating PC dimensions that can be phonetically interpreted. The idea we explore for our present dataset is to identify, based on a 'training' or reference data

[2] For reasons probably to do with individual speaker's anatomical and tissue composition characteristics, not every speaker images equally well in ultrasound. Particularly for bunched palatal or velar articulations, there may be for some speakers insufficient ultrasound reflections for image reconstruction to render anything but grey noise. See Stone (2005) for discussion.

set, for each subject a single principal component which captures the major /i, a/ related variation in tongue shape (/i/ and /a/ being assumed to represent the extremes of the vowel space in our corpus). The eigenvector corresponding to this PC is then used to compute the PC score for all frames from all utterances. The end product is a time series of a single value representing tongue shape dynamics in a subject-specific /i, a/ reference space.

We proceeded as follows: Using the acoustic segmentation information all ultrasound frames from the midpoint of V1 to the midpoint of V2 were first extracted from the /abi/ and /iba/ utterances. It was assumed that this selection of frames would provide a representative selection of tongue configurations from /i/ to /a/ with negligible influence of consonantal articulation (since these sequences did not contain a lingual consonant). The number of frames extracted in this way per subject is given in Table 2. Differences can largely be attributed to different speech rates between subjects, and in the case of S1 the lower frame-rate, related in turn to the higher scanline density.

Prior to entering the data into the PCA, a region of interest was defined individually for each subject in order to eliminate image regions with no obvious articulation-related modulation. Essentially, this cropped the data matrix of each image to eliminate pixels less than 10 mm from the probe surface and pixels more than 5 mm beyond the most extreme radial position reached anywhere by the tongue surface (in addition, scanlines at the edge of the image showing strong shadowing by hyoid or mandible were eliminated; for most subjects this involved very few scanlines). The resulting number of pixels per frame used was about 18,000 (details in Table 2).

Principal Component analysis was carried out using a singular value decomposition (SVD) approach, with centering of the data but without scaling (corresponding to use of the covariance rather than the correlation matrix). Unlike calculations based on extraction of the eigenvalues and eigenvectors directly from the covariance matrix of size $n \times n$ (e.g. for S4 size would be 17,500 × 17,500), the SVD approach allows calculations to be restricted to only those eigenvalues that can be non-zero (here $m-1$, i.e. 238 for S4). In applications of the type presented here, where $n \gg m$, such an approach is crucial to keeping memory and computation time within tractable limits (in more typical uses of PCA where the number of observations is greater than the number of variables, then the maximum number of components extracted will always be limited by the number of variables).[3] With this approach, even on a very ordinary desktop computer the computation time for one subject in the present study is only about 4 s.

In addition to showing the amount of input data for each subject, Table 2 also shows the amount of variance explained by the first three principal components. The point here is that for every subject the first component explains substantially more of the variance than the second, so there is little possibility for ambiguity as to which component captures the basic range of vowel variation from low back to high front (S4 is the only subject for whom the variance explained by PC1 is less than double that of PC2, and the only subject for whom the sum of the variance of PC2 and PC3 gets close to that of PC1 alone). Thus we assume that for all subjects the score on the first principal component for any given frame captures the configuration of the tongue in the space of vocalic articulations from /a/ to /i/. A more explicit justification of this assumption and possible limitations of using just a single axis of variation are given below.

The final data processing task prior to actual articulatory analysis was to use the eigenvector corresponding to the first principal component to compute the principal component score for all frames from all utterances. Mathematically this is the very simple operation of computing a weighted sum of the pixels in each frame. The result for every utterance is a scalar-valued function of time that is amenable to the kind of analysis and display techniques familiar for flesh-point data such as articulography (EMA) and x-ray microbeam (peak picking, velocity profile segmentation, etc., e.g., Bombien and Hoole, 2013, and many others). Before analysis, the principal component tracks were smoothed with a low-pass filter (Kaiser window, cut-off frequency 20 Hz), as is customary for instance for articulography data. Filtering in the temporal frequency domain is a first example of an operation that is trivial to apply to the output of the principal component analysis, but which is less well-defined and computationally much more costly if it has to be applied to the raw image data. A different kind of filtering, not used in the present paper, is filtering in the spatial domain to reduce image noise (e.g. the anisotropic diffusion filter for speckle reduction used by Hueber et al., 2007). Such noise-reduction applied to the raw images can be expected to increase the amount of variance explained by the first few principal components. However, the amount of variance explained in absolute terms is relatively immaterial in the present context, where the design has been structured such that the first extracted component is dominant and interpretable. Indeed, observations made in Hueber et al. (2007), and confirmed by our own

[3] Depending on the statistics package used, it may be necessary to check what the default behavior of the PCA algorithm is: For example, the Matlab function pca uses defaults that are appropriate for this application, whereas the earlier function princomp requires an explicit setting (referred to as “economy” mode) even when the SVD approach is used.

more informal observations, indicate that the patterns associated with the first few principal components are fairly insensitive to such pre-processing for noise-reduction.

3. Results

We first show the tongue configuration pattern captured by the first principal component. Before turning to the main body of the analyses we consider to what extent the interpretability of the analyses may be affected by using only the single articulatory dimension captured by this first principal component. As a precursor to the actual coarticulatory analyses we then introduce the use of the principal component scores to capture time-varying tongue shape in selected VCV sequences. We then demonstrate how the principal component scores can be extracted at any desired time-points and further processed to quantify the coarticulatory patterns in the data.

3.1. Principal component tongue shapes

Recall that the data for the principal component model for each subject used the image frames taken from the midpoint of V1 to the midpoint of V2 in the /iba, abi/ sequences. The two panels of the left column of Fig. 3 show

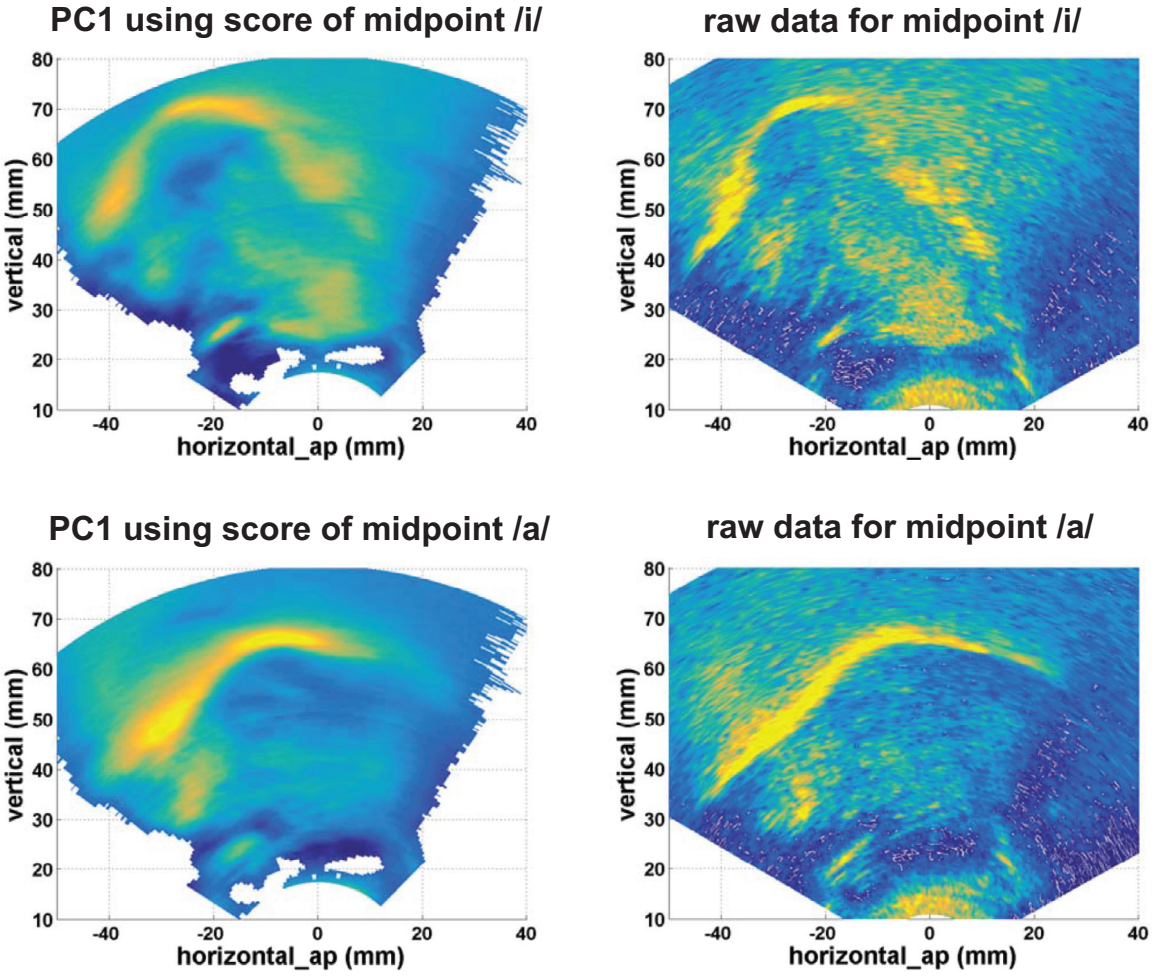

Fig. 3. Illustration of tongue shapes captured by the first principal component. The right column shows a selected representative raw data frame for /i/ and /a/ of one speaker (S7).The matching images in the left column are the image predicted using PC1 only, given the score on PC1 of the adjacent raw image. Anterior positions on the *x*-axis are to the left. See text for computational details. Differences in white background areas between left and right panels are due to (1) thresholding with respect to pixel standard deviation, and (2) cropping before carrying out PCA.

for one speaker the range of image patterns associated with the first principal component (corresponding to an articulatory nomogram in the sense used by Badin et al., 2002). These were generated in the following way: For this example speaker, typical scores on PC1 were determined for the midpoint of /a/ and /i/ in V1 position and were found to be approximately ±2100 for /a/ and /i/ respectively (the unit of the PC score is related to the variance of the raw pixel intensities, but is not otherwise informative). Two patterns of predicted pixel intensities are then calculated by (1) multiplying the pixel loadings defined in the eigenvector of PC1 by +2100 and −2100 in turn, and (2) adding back in the average pixel intensities subtracted from the raw data by the centering procedure in the PCA algorithm. This gives the pattern of pixel intensities associated with large positive and negative scores on PC1. These patterns can then be displayed in exactly the same way as raw images. Two corresponding raw images are shown in the right column of Fig. 3, i.e. two representative images whose scores on PC1 correspond to the values given above.

In the raw data frames there are substantial differences between /i/ and /a/ below the tongue surface in the body of the tongue. These differences are clearly also captured in the principal component shapes. The white background areas differ somewhat between corresponding left and right panels. This is firstly because of the cropping outlined above in Section 2.3, and secondly because in addition further pixels showing negligible variation over the training corpus were discarded before entering the data into the analysis (this was based on a threshold for pixel standard deviation).

3.2. Space of the first two principal components

We claimed above that it is plausible, given the selection of training material used, that the first principal component would capture the /i/-/a/ axis of variation. In this section we show more explicitly that this is the case, and also consider to what extent it is justifiable to leave out further principal components from the analyses carried out below. To this end we first plot selected items from the corpus in the space of the first two principal components, namely for each subject the average principal component score at the midpoints of V1, C, and V2 in the sequences /iba/, /abi/, and /ebə/ (as noted above, V1 can be regarded as the prosodically stronger position given the default trochaic German stress pattern).

With respect to PC1 it is clear in Fig. 4 that for all subjects this dimension ranges from /a/ to /i/. Moreover, in almost all cases V1 (red lower-case vowel symbols) occupies a more peripheral location than V2 (black upper-case). These effects will be quantified in detail for a wider range of items below. Schwa generally occupies a location quite close to zero. The consonants also occupy intermediate locations with respect to PC1, the precise location reflecting the temporal evolution of the V-V movement (also analyzed in the coarticulatory analyses below). Overall the picture with respect to PC1 is quite consistent over all subjects. However, the question of tongue configurations with PC1 scores that are a long way from the peripheral values leads to an important issue in interpreting the results that will first be discussed by extending consideration to PC2 and then approached in more general terms below. Based on the explanation above with respect to Fig. 3 of how the representative tongue shapes for extreme values of PC1 were generated then it follows that the predicted pixel pattern when the PC1 score is zero is simply the average value of the pixels for the images in the dataset (if only PC1 is considered). Clearly this will be an inaccurate prediction if there are additional sources of systematic variation in the data. An extreme case would be a vowel such as /u/, that is clearly similar to neither /i/ nor /a/, and thus would have a rather neutral value on PC1, but would clearly also not be similar to the average pixel pattern. This of course is the reason why in the present investigation we left such vowels out of the corpus, since it would be clear a priori that more than one component would be necessary to model the data. The question nonetheless remains whether even the present corpus shows systematic additional sources of variation. A first answer can be given to this by considering in Fig. 4 whether PC2 also shows a consistent pattern and, if so, what its articulatory interpretation might be. Two simple points to be made straightaway are (1) that PC2 not surprisingly makes no contribution (except very marginally for S1) to the /a/-/i/ distinction, and (2) that it is indeed much more difficult for PC2 than for PC1 to make general statements that are valid for all 7 subjects. S1 and S4, in particular, are markedly different from the other subjects (and from each other). But this means that we can try to summarize common features for the remaining subjects S2, S3, S5, S6 and S7. For these subjects PC2 seems to capture the difference between peripheral and intermediate tongue configurations, i.e. it has large positive values for both large positive and negative values of PC1, and large negative values when PC1 is close to zero, i.e. for schwa and usually one or more of the consonantal items. In a sense, then, it captures details of how tongue configurations half-way between /a/ and /i/ with respect to PC1 differ from the overall average pixel pattern. Generating the tongue

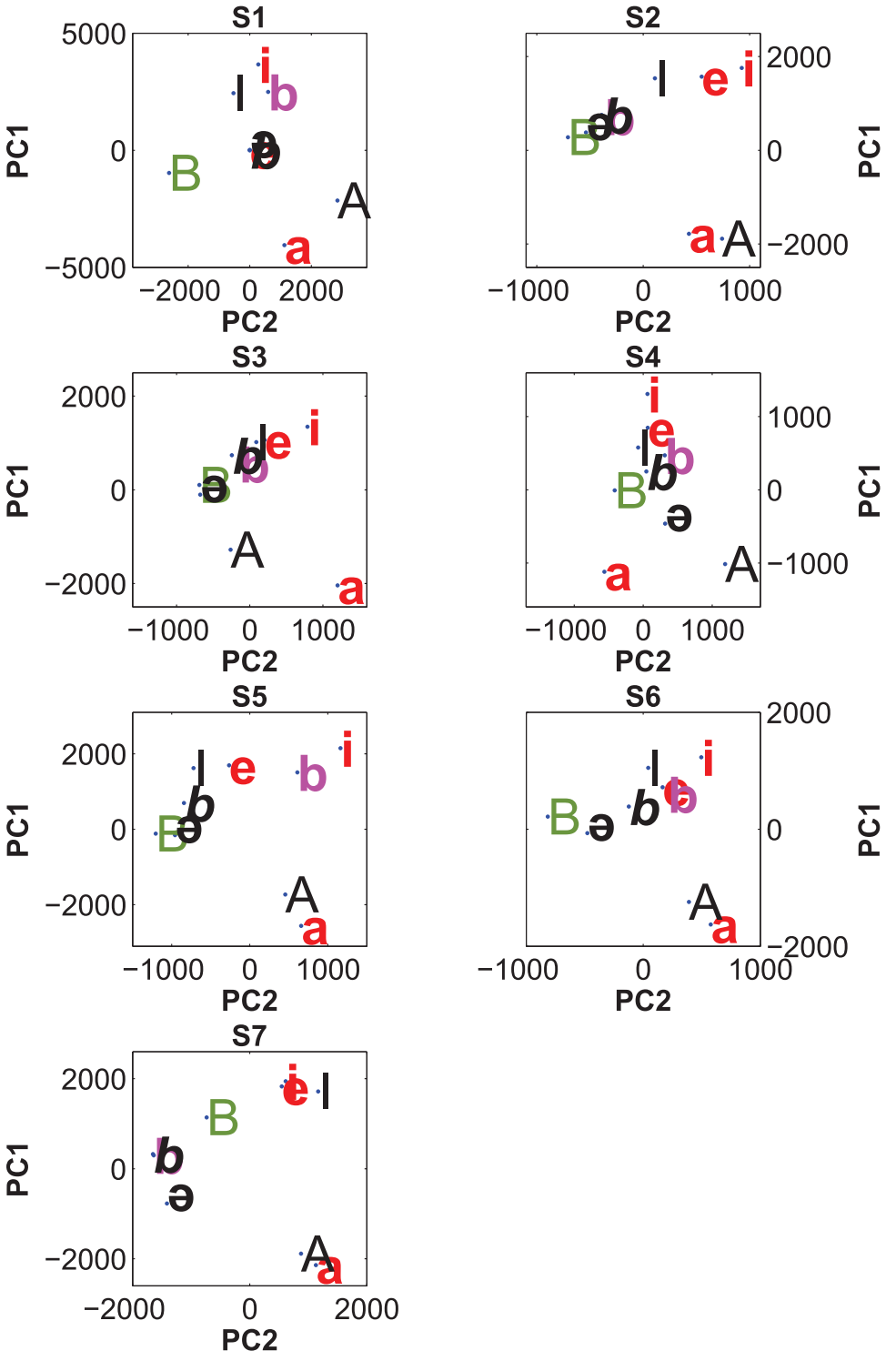

Fig. 4. Principal component scores extracted at the midpoint of V1, C, and V2 in /iba/, /abi/, and /ebə / (averaged over all repetitions) and plotted in the space defined by the first two principal components. Separate panel for each speaker. Coding: lower-case red for V1 (/a, i, e/); upper-case black for V2 (/a, ɪ/; also black for V2 = / ə /); /b/ from /abi/ upper-case dark-green; /b/ from /iba/ normal font magenta; /b/ from /ebə/ italic font black. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

shape as in Fig. 3 for values of PC2 at the schwa end of this dimension indeed showed a tongue contour intermediate in location between /a/ and /i/. Thus while there is something to be said for basing further articulatory analyses on a space of more than one dimension this would be extremely awkward in practice, because as just noted, two speakers do not fit in with the PC2 pattern just discussed, and even with the remaining five there are further considerable differences of detail, for example whether and how PC2 captures V1-V2 differences for /i/ and /a/. The degree of distortion through data reduction can be assessed in the following way (see also Fig. 3). Essentially, for every sound in the corpus we can use its score on PC1 to generate the pixel pattern predicted on the basis of PC1 alone. This predicted pattern can then be subtracted from the measured pattern (i.e. right panel minus left panel in Fig. 3) and a measure of reconstruction error can be defined as the RMS value over all pixels of this difference. Interpretation based only on PC1 should be done cautiously for sounds that have high reconstruction error values. Fig. 5 shows the results for V1 and V2 based on average values for each speaker. The figure is based on the five speakers (i.e. S1 and S5 are omitted) who also recorded a small amount of additional material, which happened to include two tokens of the vowel /u/. This is included in the figure for comparative purposes and indeed as expected shows the highest values for reconstruction error.

The other vowels fall into two groups. A low level of reconstruction error is found for /i/ (equally low in both V1 and V2 position), and for /e/. This is particularly notable for /e/ since it indicates that the score on PC1 provides just as good a model of the tongue configuration as it does for /i/, even though /e/ did not form part of the material used to set up the model (note that reconstruction error will never be zero, simply because of noise in the image, as is very visible in Fig. 3). This is maybe not that surprising given the articulatory proximity of /i/ and /e/ which are both articulated with a strongly bunched tongue shape. /a/ and schwa show an intermediate level of reconstruction error compared to /i, e/ on the one hand and /u/ on the other (for /a/ this applies to both V1 and V2 position). For schwa this is immediately understandable from the above discussion of PC2, i.e. from the fact that for many of the speakers schwa is located at one of the extremes of the PC2 dimension of variation. The overall higher reconstruction error for /a/ than /i/ is less easy to explain. We suspect it may be due to a somewhat greater difference between V1 and V2

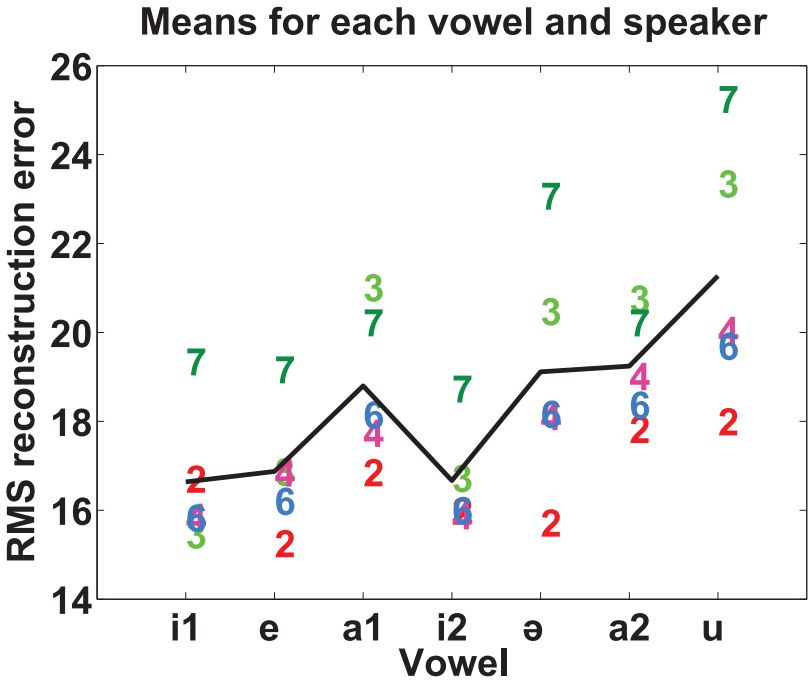

Fig. 5. Reconstruction error defined as RMS difference between original image and image reconstructed from score on PC1 only. Computed for each vowel in the corpus separately for V1 and V2 position (vowel and position indicated by the *x*-axis labels; /e/ and /ə/ occur only as V1 and V2, respectively). Data points for /u/ (far right) are from additional material not in the main part of the corpus. Each data point (labeled by speaker number) is the average for one speaker of all tokens with the given vowel, based on the image frame at vowel mid-point (speakers also color-coded). The black solid line shows the mean over all speakers. Units are arbitrary (determined by the storage format of the raw pixel data, i.e. 8 bit integers). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for /a/ than /i/, so that the /a/ end of the PC1 dimension is some kind of compromise pattern between V1 and V2. However, if there are systematic differences between /a/ in V1 and V2 position that are not captured by PC1 then Fig. 4 makes it clear that they are not captured by PC2 either. So to throw more light on this issue it would be necessary to look in detail for each speaker at further principal components (each explaining rather little variance). While this would be beyond the scope of the present paper this discussion overall serves to highlight the basic problem associated with PCA generally, which is mapping the PC scores onto meaningful parameters. This means that the feasibility of PC analysis for image data may greatly depend on the specific speech material, a point we shall return to in the Discussion.

We now turn to the principal component scores of all consonant and vowel contexts, first demonstrating for selected VCV sequences how our method captures time-varying tongue shape in this /i, a/ space defined on the basis of the training set.

3.3. Principal component scores as time functions

We first illustrate the results for the different consonants in the /i, a/ contexts and then turn in the next section to the results for the additional vowel contexts /e, ə/. Figs. 6 and 7 show, for one subject each, the time-varying patterns of the first principal component scores for the consonants in our corpus (individual panels for /b, d, l, n, ʃ, z /) and 4 vowel contexts (different colored lines within each panel for /iCi, aCa, iCa, aCi/). The data are lined-up (zero on time axis) at the acoustically determined onset of the consonant. The circles overlaid on each line correspond (from left to right) to onset of V1, onset of consonant (= offset of V1), offset of consonant (= onset of V2), offset of V2. Each curve is an ensemble average over (typically) 7 repetitions of each VCV sequence. For ensemble average calculation the individual tokens were first time-warped to the average duration of V1, C, and V2 for the relevant sequence.

Before making some general observations based on these figures (that will then be looked at quantitatively and in more detail in the quantitative analysis sections below) it is worth emphasizing how useful it is, in terms of obtaining a quick overview over major patterns in the data, to be able to average all utterances of a given type and then overlay graphically different utterance types / conditions with a free choice of lineup points.[4] Such operations would be quite complicated to define for tongue contour data (cf. Fig. 1 right), quite apart from the issues associated with reducing a tongue contour to a single number: interpolation and averaging of contours assumes that the contour can be captured as a fixed number of points (e.g. 100) that can be considered homologous across frames. This is a notoriously problematic assumption if varying portions of the tongue surface are obscured by the hyoid or mandible acoustic shadow. Also variation in tongue length between articulations due to compression and stretching as typical for hydrostatic motion is a severe issue for the assumption of point homology, as is the fact that the ultrasound image may capture different parts of the vocal tract for different speakers. Since no anatomical reference structures are visible in ultrasound, it is not really possible to calibrate the region of the vocal tract captured across recording sessions and speakers. The PC analysis circumvents this problem by enabling us to exploit whatever systematic information is present in the image for setting up the relative /i, a/ space for each speaker. This of course hinges on the availability of a condition (here: the labial) which justifies a high degree of confidence in the interpretation of the first PC as /i, a/ space for each speaker, as discussed in detail in the previous section.

For a first approximation to the results, a visual indication of anticipatory V-to-V coarticulation in Figs. 6 and 7 is given by looking at, for example, the extent to which the /aCi/ curve deviates from the /aCa/ curve at 0 on the time axis (onset of the consonant). This gives a measure of the anticipatory influence of /i/ on /a/. Similarly, the deviation of the /iCa/ curve from the /iCi/ curve at the line-up point gives the anticipatory effect of /a/ on /i/. Considering this over all consonants (and subjects) gives a measure of whether /a/ or /i/ as V2 exerts a stronger influence on V1 (a rough visual indication of /i/ exerting stronger effects would correspond to the /iba/ and /abi/ curves intersecting closer to /i/ on the *y*-axis). In addition, one can ask whether the strength of these V-to-V effects also depends on the intervening consonant as described by Öhman (1966). The /b/ context can serve as the reference condition in which maximal vowel-to-vowel coarticulation can a priori be expected (since the labial does not control tongue shape).

[4] Ensemble-average tracks of this kind could also be plotted with error-bars showing the amount of variation at each time-point. However, since this would potentially distract here from the main aim of giving a graphical illustration of coarticulatory phenomena, we will look briefly at the issue of token-to-token variability with data from all subjects in Section 3.4 (Table 3b) below.

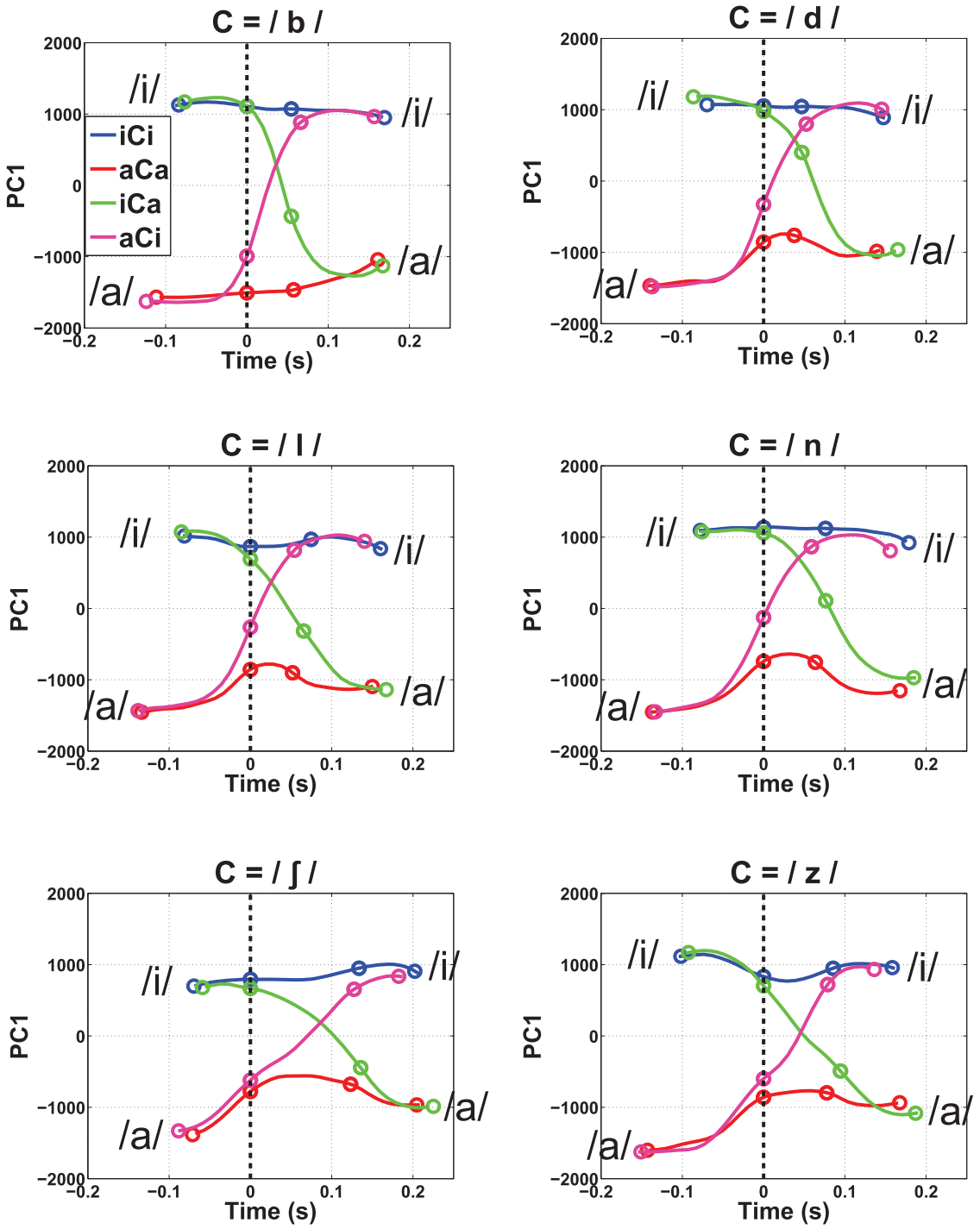

Fig. 6. Principal component scores as a function of time for selected VCV sequences of speaker S6. In each panel C is constant. Color-coding of V1 and V2 combination is shown in top left panel. Lineup (shown by vertical dashed line at $t=0$) is V1 offset in all cases. See text for details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The corresponding procedure for carryover articulation can be carried out by aligning the curves at consonant offset and then measuring the deviation between /aCa/ and /iCa/ (and /iCi/ and /aCi/) separately for each consonant. This would then give overall a measure of whether anticipatory or carryover coarticulation is stronger. For reasons

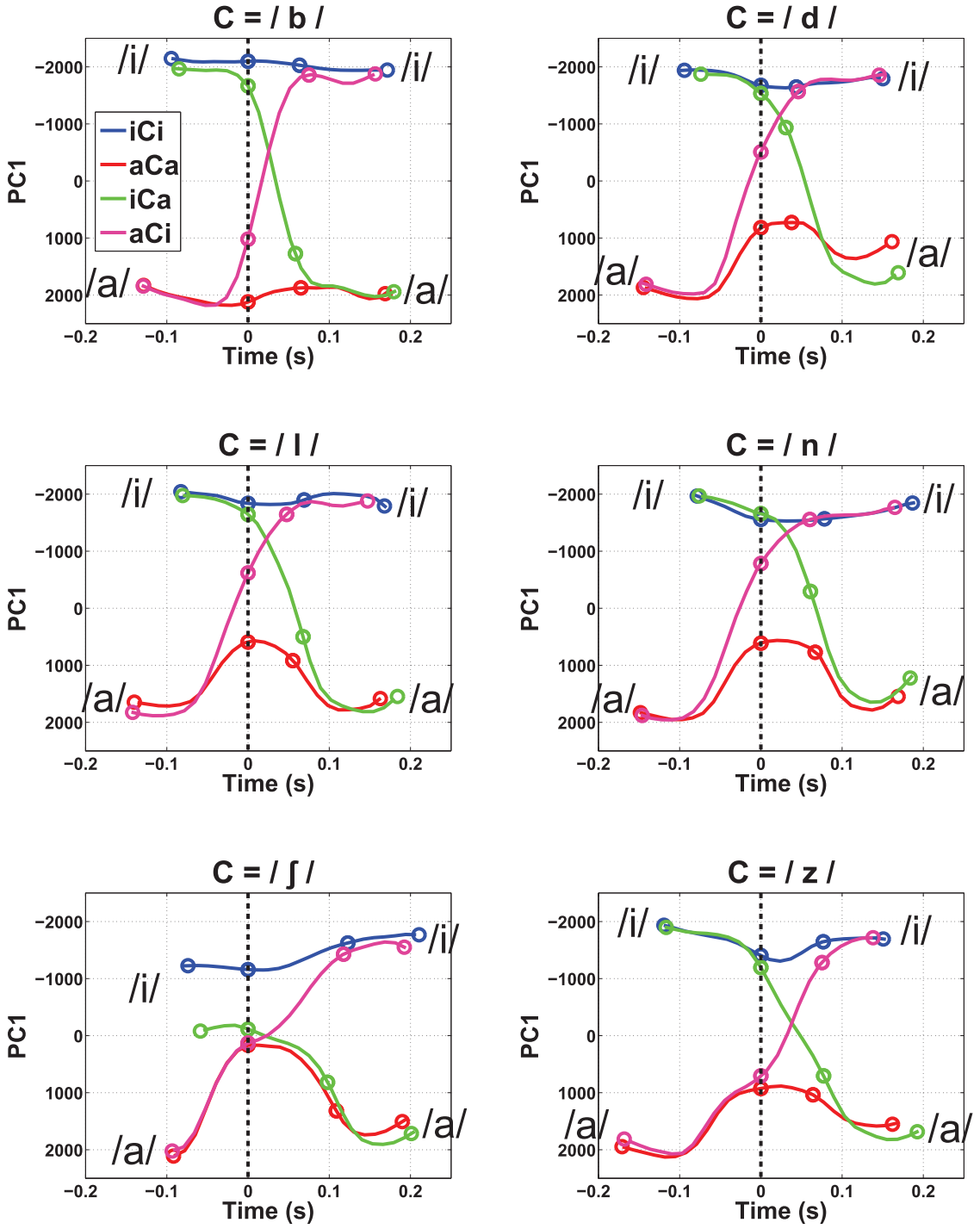

Fig. 7. Principal component scores as a function of time for selected VCV sequences of speaker S7. Further details as in Fig. 6.

of space we do not show separate figures with consonant offset as line-up point. However, by way of example, a reasonable appreciation of the relative strength of anticipatory and carryover effects of /i/ on /a/ can be gleaned from comparisons such as the following: Compare the deviation between the /aCi/ (magenta) and /aCa/ (red) curves at the line-up point (dashed line at $t=0$ in Figs. 6 and 7) with the deviation between the /iCa/ (green) and /aCa/ curves at

the first marked point following the line-up point (i.e. the consonant offset, usually located between 0 s and 0.1 s on the time axis). If the latter deviation is the larger one this indicates a stronger carryover effect (this is the case, for example, in the top panels of Fig 6, for C=/b/ and C=/d/).

In addition to V-to-V coarticulation, it is also possible to consider the extent to which the vowels are affected by the different consonants. The examples in Figs. 6 and 7 illustrate several cases in which the tongue configuration at the midpoint of V1 (and indeed sometimes even at the onset of V1) differs between /VbV/ sequences and corresponding cases with lingual consonants. A particularly striking example is /ʃ/ for S7, perhaps not surprisingly given that /ʃ/ can be assumed to show more dorsal involvement than the other lingual consonants in the corpus.

We now flesh out some of the qualitative observations made with respect to Fig. 4 about the representation of the vowels in the principal component space to consider quantitatively how the /i, a/ axis of variation, set up based on the training set, captures the other two vowels included in the study.

3.4. Quantitative analysis of basic vowel patterns

Before turning to a complete numeric breakdown of coarticulatory effects of the kind just outlined above we will illustrate the use of the PCA scores for capturing very basic vocalic configurations, i.e. /i, e, a/ for V1 and /i, a, ə/ for V2. For a straightforward combination of results over subjects the data were normalized for each subject such that the PC value for the midpoint of V1=/i/ in /ibi/ was mapped to +1 and V1=/a/ in /aba/ to −1. The results for each vowel category, averaging over subjects and all contexts but broken down by syllable position, are given in Table 3a (the /ʃ/ consonant context has been left out because of its strikingly large effects on vowel midpoint data, visible in Figs. 6 and 7, and analyzed quantitatively in Section 3.7 below). Calculations were based on the ensemble-averaged data, so each of the 100 data points for /a/, for example, was in turn the average of approx. 7 repetitions. Data was extracted at the temporal midpoint of the vowel.

Table 3b gives a similar breakdown of token-to-token variability, i.e. proceeds exactly as for the data in Table 3a except that the standard deviation rather than the mean over the 7 contours in each ensemble is calculated.

Table 3a confirms first of all the impression gained on the basis of Fig. 4 that the PCA gives phonetically interpretable results for data that was not in the training corpus: Values for /e/ are clearly positive like /i/, but slightly smaller, reflecting a somewhat less extreme palatal articulation. Values for schwa are very close to zero, meaning roughly halfway between /i/ and /a/. In addition, both /i/ and /a/ evidently show less extreme values for V2 than for V1, almost certainly reflecting the weaker stress on the second syllable (this is also visible to a certain extent in Figs. 6 and 7 above, especially for /a/). In addition, variability is a good deal higher for vowels in the second syllable, probably reflecting some variation in the amount of reduction over speakers and contexts for the more weakly

Table 3a
mean values (±standard deviation) of principal component scores at vowel midpoint.

	V1	V2
a ($n=100$)	−0.95±0.100	−0.70 ±0.197
i ($n=70$)	+0.96 ±0.060	+ 0.81±0.182
e ($n=60$)	+0.79 ±0.105	
ə ($n=60$)		−0.06 ±0.294

Table 3b
token-to-token variability based on the mean (±standard deviation) of the standard deviation of principal component scores at vowel midpoint in each ensemble.

	V1	V2
a ($n=100$)	0.10±0.038	0.14±0.080
i ($n=70$)	0.07±0.027	0.09±0.047
e ($n=60$)	0.10±0.065	
ə ($n=60$)		0.14±0.063

stressed syllable. This variability is apparent both with respect to the mean tongue configurations given in part (a) of the table as well as with respect to the token-to-token variability given in part (b). In addition, it will be observed from part (b) that token-to-token variability appears to be lowest in /i/, which may be consistent with suggestions that /i/, as a high palatal vowel, can exploit bracing against the hard palate to achieve stability.

In sum, our results so far demonstrate that defining a reference /i, a/ vowel space using the first PC can be used as the basis to generate time series of PC scores that can easily be subject to standard numeric operations such as averaging and time warping. Moreover, the PC values obtained are clearly interpretable in terms of the expected coarticulatory effects.

3.5. Quantitative analysis of coarticulation in schwa

The analysis underlying the baseline values for overall vowel configuration just shown above can be extended in a direct way to a first very typical coarticulatory question: To what extent is the tongue position in schwa systematically influenced by V1? For this, data was available from the sequences /eCə/ and /aCə/ (but no data for S1, see Section 2). The value for schwa in the V1 = /a/ context was subtracted from the value in the /e/-context for each matched pair of subject and consonant (here and elsewhere in this paper data was first averaged over repetitions, i.e.. the data for further analysis was extracted from the ensemble-averaged data).

The left panel of Fig. 8 shows examples of two relevant VCV sequences (subject S4). The left-most markers overlaid on the time-functions correspond to V1-offset (=onset of /n/), the middle markers (at $t = 0$) to schwa-onset (= offset of /n/) and the right-most markers to offset of schwa. These examples illustrate the expected coarticulatory pattern that at the midpoint of schwa (around $t = 0.025$) the tongue position is perturbed towards the neighboring V1.

The full set of difference values for all subjects and consonant contexts are plotted in the right panel of Fig. 8. If there were no systematic influence of V1 on the schwa the data would cluster around zero. The expected coarticulatory pattern would generally be that the schwa is shifted towards higher values in the /e/ context and lower values in the /a/ context, giving positive values for the difference calculation. This is indeed the predominant pattern in the right-hand graph of Fig. 8: the mean over the 36 data points in the figure is 0.13 (sd = 0.21). Note, however, two main exceptions: firstly with respect to the speakers, the values of S2 were consistently slightly negative, indicating that global tongue position during schwa in the /a/-context was actually slightly more i-like than in the /e/-context. On the other hand, this overview of all subjects indicates that S4, used for the illustration in the left panel, is quite a typical subject. With respect to the consonants, /ʃ/ displays no clear trend away from zero. This is not unexpected:

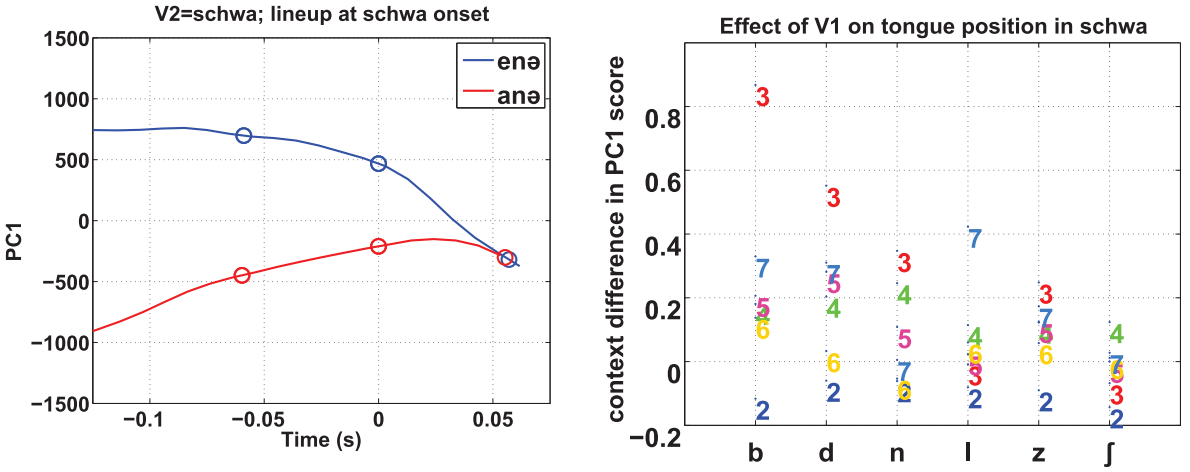

Fig. 8. Coarticulatory patterns in schwa. Left panel: Representative time functions for one speaker (S4); /e/-context in blue, /a/-context in red. The left-most markers overlaid on the time-functions correspond to offset of V1 (=onset of /n/), the middle markers (at $t = 0$) to onset of schwa (= offset of /n/) and the right-most markers to offset of schwa. Right panel: Summary over all subjects and consonant contexts of the difference in PC1 score for schwa in /e/-context vs. /a/-context. Data points are labeled with speaker number (speakers also distinguished by color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Since /ʃ/ is known to be least susceptible to the influence of surrounding vowels (Stone et al., 1992), it will also be least transparent for vowel-to-vowel coarticulation. Also in line with Stone et al.'s results, /z/, by contrast, is more amenable to coarticulatory permeability. Interestingly, and consistent with Proctor's (2011) claim, /l/ blocked carryover vowel-to-vowel coarticulation almost as much as /ʃ/, whereas /d/ patterned rather with /b/ as far as coarticulatory transparency is concerned. Although S2 deviated from the overall trend, the difference between tongue configuration in the two V1-contexts was still significant in a paired *t*-test ($t[35] = 3.8$, $p < 0.001$).

3.6. Quantitative analysis of V-to-V coarticulation

We now consider a more detailed analysis of V-to-V coarticulation (refer back to Figs. 6 and 7 for graphical illustration of the procedure below). A normalized measure for the amount of anticipatory and carryover coarticulation in /i/ and /a/ was calculated as follows: To measure anticipatory coarticulation data was extracted at the end of V1 for the VCV sequences defined in the following formula, where subscript indices refer to vowels in the set V = {i, a}. The *coarticulation score* AC is then given by

$$\begin{aligned} &AC_k = \left(V_kCV_k - V_kCV_j\right) / \left(V_kCV_k - V_jCV_j\right) \\ &k = 1,\ 2;\ j = 3-k \end{aligned} \tag{1}$$

C is set in turn to each consonant in the corpus. For carryover coarticulation the same calculations were carried out, except that data was extracted at the onset of V2.

The strength of coarticulation is given by a value between 0 and 1. For example, if at the end of V1 /i/ in /ida/ is identical to /i/ in /idi/ there is no anticipatory influence of /a/ on /i/ (AC(/i/) = 0). If the tongue position is half-way between /i/ in /idi/ and /a/ in /ada/ then AC(/i/) = 0.5.

Since this procedure was carried out for all combinations of speakers and consonantal contexts, the normalization by means of the symmetric sequences /iCi/ and /aCa/ (in the denominator of the above equation) was consonant-specific.

Before presenting the results we would like to introduce a technical note: Since complete tongue configurations are represented by a single number, it is very easy to interpolate values for arbitrary time-points. This was done here (simple linear interpolation was deemed sufficient, given the generally quite high frame-rate of the ultrasound data): the acoustically defined time-point of the offset of V1, for example, does not necessarily correspond to the time at which a specific ultrasound frame was acquired. And, of course, the above analysis could be expanded to multiple time-points with virtually no additional work.

With the above procedure, matching coarticulation scores for /i/ and /a/ were available for each combination of speaker and consonant. Accordingly, we will focus first on the matched differences between /a/ and /i/ coarticulation scores, before looking at the overall magnitude of anticipatory and carryover effects.

Fig. 9 shows delta scores which were computed by subtracting the /a, i/ coarticulation scores from each other such that positive values correspond to a greater coarticulation score for /a/ than /i/. A value of zero indicates that both vowels influence each other to the same degree. The data are arranged on the abscissa by consonantal context, with individual data points plotted for each speaker. The general pattern to be expected here based on previous research (e.g., Recasens et al., 1997) is that /i/ should exert a greater influence on /a/ than vice versa (have more coarticulatory 'aggression'). This means we expect positive delta scores in terms of our measure (meaning /a/ is more susceptible to influence from a preceding or following /i/ than vice versa). Again we may ask at the same time whether this is true for all consonants.

For both anticipatory and carryover effects the majority of values were positive, i.e. /a/ was more prone to coarticulatory influence than /i/. However, this trend was much clearer for consonants /b, d, n/ and to some extent /l/, whereas values for the fricatives /z, ʃ/ clustered around zero (with one very large negative value, for /ʃ/ of speaker 7). These observations are summarized in Table 4, showing that the average coarticulation value is more strongly positive (and the standard deviation is smaller) when the fricatives are left out of the calculation.

For anticipatory coarticulation the difference between /a/ and /i/ was significant ($t[40] = 4.3$, $p < 0.0001$) in a paired *t*-test, even without eliminating the fricatives. For carryover effects the difference was significant at $p < 0.01$ ($t[40] = 3.4$) using all consonants, but at $p < 0.0001$ ($t[27] = 5.4$) if the fricatives are left out.

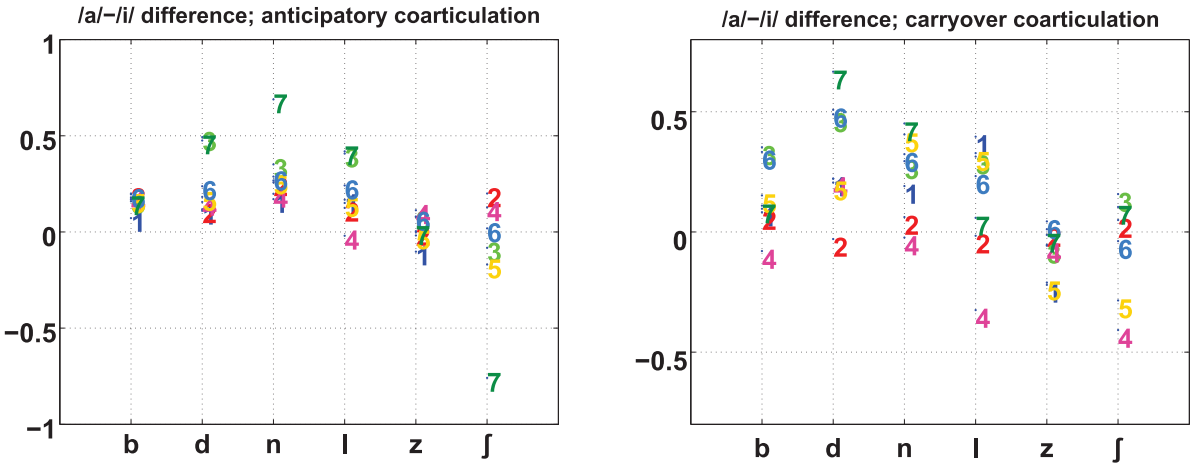

Fig. 9. Coarticulatory differences between /a/ and /i/ (positive values indicate a stronger coarticulatory effect in /a/ (greater coarticulatory aggression of /i/ compared to /a/). Data points are labeled with speaker number (speakers also distinguished by color). Anticipatory effects in the left panel, carryover in the right panel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Difference in coarticulation values of /a/ and /i/ (mean±sd).

	Anticipatory	Carryover
All consonants ($n=41$)	0.12±0.237	0.15±0.221
Without fricatives ($n=28$)	0.24±0.148	0.22±0.211

Based on this data we can also look more closely at the relative magnitude of carryover and anticipatory coarticulation. Fig. 10 plots the difference with respect to coarticulation direction for data matched by speaker, consonant context (and vowel): for each condition the value of carryover minus anticipatory is plotted, i.e. values are positive if carryover effects are stronger (data for /a/ in left panel, for /i/ in right panel).

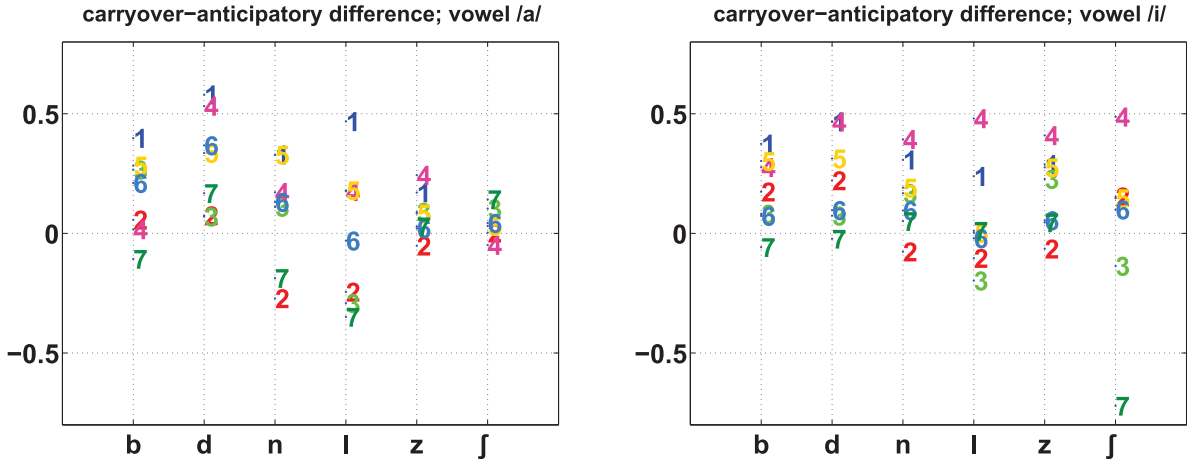

Fig. 10. Difference between carryover and anticipatory coarticulation effects (positive values indicate stronger carryover effects), shown separately for /a/ and /i/. Data points are labeled with speaker number (speakers also distinguished by color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

These data indicate a clear trend for carryover to be stronger than anticipatory coarticulation. The unusual value for /ʃ/ of speaker 7 is once again visible as the only case where anticipatory effects were much stronger; however it is here much less clear whether fricatives generally pattern differently from the other consonants. Means and standard deviations over all consonants were 0.11 ±0.214 for the matching items with /a/ as vowel, and 0.14 ±0.227 for /i/ (i.e. over the data in the left and right panels of Fig. 10, respectively). In a paired *t*-test (with data aggregated as in Fig. 10 over all combinations of subject and consonant) carryover effects were stronger for /a/ ($t[40] = 3.4$, $p < 0.01$) and for /i/ ($t[40] = 3.9$, $p < 0.001$). The latter case was thus clearly significant even without removing the /ʃ/ items.

3.7. Quantitative analysis of C-to-V coarticulation

Finally, as an example of C-to-V coarticulation we consider the extent to which the tongue configuration at the midpoint of the vowel is affected by the adjacent consonant. As mentioned above, we used VCV sequences with consonant /b/ as baseline and compared sequences with the lingual consonants to this, each comparison consisting of a pair of VCV sequences with identical vowels (e.g. /ida/ was compared with /iba/ at the midpoint of V1 and V2). Typically, the effect of the lingual consonant was to shift the tongue configuration to a less extreme position compared to the bilabial consonant context, thus effectively showing heightened influence of the consonant. Thus subtracting the V1 value of /ida/ from the V1 value of /iba/ would give a positive value if the /i/ in /iba/ had a more strongly positive PC1 score than in /ida/. Conversely, subtracting the V2 value of /ida/ from /iba/ would give a negative value if the PC1 score for /a/ in /iba/ was more strongly negative.

We restricted the analysis to sequences with /i/ and /a/ as this gave a more balanced set of data for comparing anticipatory and carryover effects. For the anticipatory effects, the comparison between labial and lingual consonant contexts was made at the midpoint of V1, and at the midpoint of V2 for carryover effects. Results for /i/ and /a/ are shown with separate lines (solid for /i/ and dashed for /a/) in the two panels of Fig. 11. Each data point shows the mean and standard deviation of the difference from the matching bilabial sequence over all subjects and contexts for the corresponding combination of vowel and consonant. Thus the value in the left panel for the dashed line (/a/) at the 'd' location on the abscissa is the mean over all subjects for the sequences /adi/ and /ada/, thus usually the mean over 14 items: 7 subjects × 2 sequences (for /ʃ/ only 6 subjects were available, see Methods).

Fig. 11 clearly shows the presence of anticipatory C-to-V effects, since values are systematically positive for /i/ and negative for /a/ (i.e. in most cases in the left panel zero is not within the range delimited by the error bars). Moreover, these effects appear to be particularly strong for /ʃ/, which, as already pointed out with the ensemble average displays, probably comes about because this consonant in our corpus has the strongest dorsal involvement.

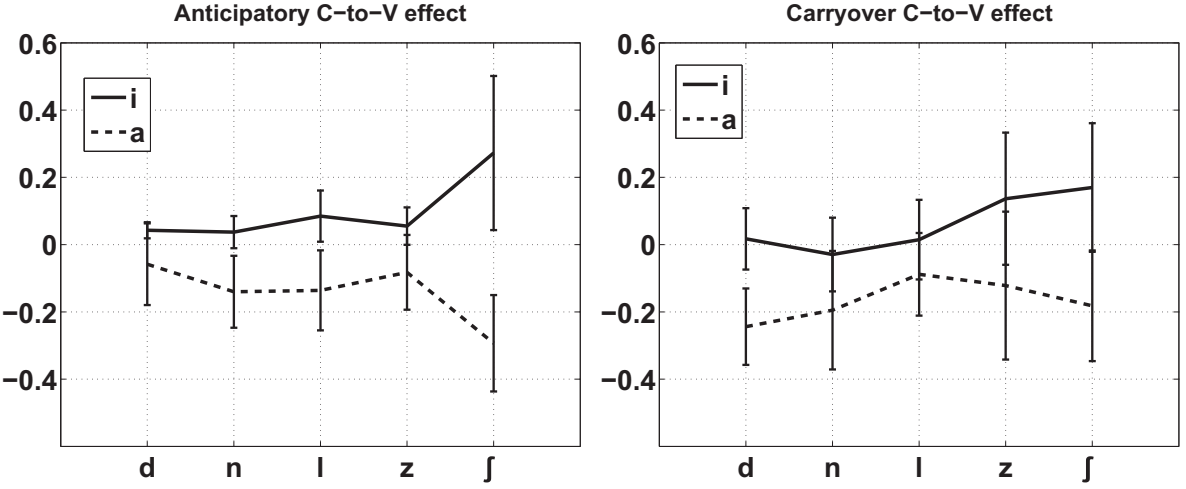

Fig. 11. Breakdown of C-to-V coarticulation effects. Left panel: Effect of C measured at the midpoint of V1. Separate lines for V1=/i/ (solid) and V1=/a/ (dashed). Right panel: Effect of C measured at the midpoint of V2. Effects are calculated by subtracting the PC1 score at vowel midpoint in /VbV/ from the matching sequence with the lingual consonants on the abscissa. Mean and sd over speakers and matching V_V contexts.

Carryover effects (right panel) show broadly the same tendency for positive values for /i/ and negative ones for /a/, but taking the error bars into account the effects appear weaker than in the left panel (i.e. are less robustly different from zero). Also there is no difference between /ʃ/ and the other consonants.

4. Discussion

It was the goal of this paper to illustrate the use of PCA as an efficient data reduction technique that can be applied to the entire ultrasound image, circumventing the current major bottleneck in ultrasound speech research which is contour tracking. Our method required a minimal amount of data pre-processing. We did for instance not apply any filtering algorithms to the images prior to processing; only the final time series of the first principal component was subject to smoothing by means of an ordinary low-pass filter. Only a single region of interest was defined per speaker. The manual intervention actually required was not for evaluating the ultrasound sequences as such but for correcting the automatic acoustic segmentation. Using this method, we were able to reduce image sequences to a single time-varying value and extract known patterns of coarticulatory variation from the data.

PCA is a non-parametric technique designed to extract patterns from high-dimensional data by finding orthogonal axes of variation in terms of eigenvectors and their associated eigenvalues. The eigenvectors are calculated from overall distributional properties of the data and independently of the variables used to construct the dataset. A well-known challenge in PCA analysis is therefore the interpretation of the resulting scores, since the major axes of variation may not straightforwardly map in a meaningful fashion onto articulatory parameters. While PC1 had in our data a very clear interpretation, for PC2 problems associated with the interpretatbility of PC scores became evident. At the same time, the dataset contained with respect to the training corpus a single main experimental variable (vowel context). That is, the fuzzy interpretability of PC2 may also be related to there being only a single defined independent variable in the design of the experiment. Obviously, the challenge of assigning meaning to the scores increases with the dimensionality of the data and the number of PCs needed to adequately capture the main patterns of variation. For ultrasound, the amount of variance captured by the first component (cf. Table 2 in Section 2) is indeed small in absolute terms compared to perhaps more familiar applications in which PCA is applied to articulatory data based on fleshpoint or x-ray gridline data. For instance Carignan et al. (2016) retained 20 PCs in their ultrasound analysis in contrast to PCA of articulography and x-ray data with just a few PCs in Hoole (1999) or Maeda (1990). While Carignan et al. find an elegant way of mapping their 20 PC scores onto formant distance values with a known articulatory correlate, our analysis retained a single PC. This was possible only because our corpus, unlike the quite varied corpus used in Carignan et al.'s study, contained a training set or control condition with one very clear and predominant axis of variation (between /a, i/). This was very advantageous for the present analysis, but at the same time is of course also a restriction. More complicated corpora for which multiple axes of variation need to be assumed may require a different approach, since a PCA of the whole tongue may not result in components with a straightforward phonetic interpretation, nor may easy comparability across speakers be guaranteed. We have not attempted in this paper to consider whether and how a single principal component model can be defined across multiple speakers, simply because for our specific purposes it is a reasonable assumption that for young healthy adults of a homogeneous variety of German the axis defined by the first principal component is comparable across speakers even when set up on a speaker-by-speaker basis. Clearly such an approach could not be applied if the aim is to compare, for example, differences between dialects, or between normal and disordered speakers. It would be an interesting topic for further research to explore whether intrinsically multimodal procedures such as Parafac (Harshman and Lundy, 1984; Hoole, 1999), where one of the modes is constituted by the speakers, can be applied to this kind of image data.

As just emphasized, our exemplary analyses exploited relatively gross vocal tract shape configurations – the /i, a/ space covers probably the most marked tongue shape differences to be found in speech sounds. At the same time, we were able to successfully capture effects beyond /i, a/ in the /e, ə/ conditions; also, the expected effects of linguistic stress on vowel production were evident in the results. Generally, many research question in phonetics trace quite subtle variation in articulator position over time, often of no more than a couple of millimeters. Also separate analyses of the (somewhat) independently controllable parts of the tongue (tip, blade, dorsum) which implement the linguistic parameters of constriction location and degree of a given sound are common in phonetic research (Bombien and Hoole, 2013; Proctor et al., 2011). The type of method applied here can in principle be extended to accommodate these requirements: For instance, depending on one's research question and corpus design it may be desirable to first define regions of interest that can be assumed to be characteristic for a particular class of articulation. This was done

for instance in Proctor et al. (2011) for MRI data. In that work, pixel intensity changes in the MRI image sequences were tracked for two separately defined dorsal and coronal regions of interest, from which closure and release phases of these two articulators could be analyzed for Italian geminates. Using the same method, Parell and Narayanan (2014) present an analysis of rather subtle coronal reduction patterns in English and Spanish. In pilot work applying PCA to MRI data we likewise achieved promising results in extracting the time-course of velum movement by applying the PCA to a region of interest in the velum region (see also Carignan et al., 2015; Fu et al., 2015).

The focus in the present paper has been to work within the framework of Öhman's basic insight of the underlying V-V movement present in VCV sequences. Thus we have not looked at specific properties of the intervening consonants. However, there has of course been much discussion of active dorsal involvement in many coronal consonants, which would be a natural target for analysis. One approach in the spirit of Öhman to extending the present work in this direction might proceed as follows: (1) set up a PCA model as here using very "vocalic" material, (2) extract a representative set of ultrasound frames from the coronal consonants, (3) based on the PC score on the vocalic component subtract out the estimated vocalic contribution to these consonant frames (basically using the computational procedures illustrated above in Fig. 3), (4) subject this non-vocalic residual in the consonantal frames to a separate PCA analysis. If combined with the approach outlined above of defining characteristic regions of interest for key constriction locations the overall approach would be in the spirit of the kind of guided principal component analysis pioneered by Maeda (1990).

A further interesting topic for further work would be to pin down in greater detail just how much information is provided by locations in the tongue close to the probe (i.e. below the air−tissue interface). In addition to simply comparing principal component models based on data extracted over a range of different depths from the probe, it would be particularly interesting to combine this with the approach of Carignan et al. (2016) in which an acoustically defined measure is used to constrain the final model. Thus it would be interesting to determine how well vowel acoustics can be predicted even if the tissue−air interface itself is left out of consideration.

There will certainly, however, remain research questions which are aimed to uncover, for instance, subtle trading relations across articulators (e.g., Smith, 2014) that are more successfully captured by vocal tract segmentation data extracted from tracking algorithms. The type of PCA analysis pursued here may thus not be applicable to all types of studies; conversely it will be important to design one's corpus with the potentials and restrictions of these novel data analysis techniques in mind. Moreover, the interpretational scope of the PCs will greatly depend on the training corpus − for instance, for our present analysis, the PCs are informative about similarity of a given frame to either /i/ or /a/. A more elaborate testing of the PCA approach to a whole variety of study designs will have to be done to gauge the range and types of research questions it may successfully be applied to.

Finally, we have emphasized that our method allows a comprehensive assessment of tongue shape *dynamics,* because the method straightforwardly lends itself to the evaluation of time series data and not just to analyses based around single time-points. Since our main objective in this paper has been to show the potential of this method, we have used very simple statistical techniques to give a basic idea of the robustness of the patterns presented. Of course, the time series of the PC scores need not be restricted, as here, to static analyses of selected time points. It is also in this respect that there have been many recent advances in statistical methods (see e.g., Cederbaum et al., 2016; Morris and Carol, 2006; Wieling et al., 2014). But in contrast to the ideas presented in the present paper, these statistical innovations currently require great computational power, certainly more than what is typically provided by an ordinary desktop computer. Yet the kind of processing applied here makes ultrasound and more generally image data in principle very suitable for these innovative statistical procedures, and increasing the latter's computational efficiency will surely only be a matter of time.

When Öhman published his seminal paper on coarticulation 50 years ago (Öhman, 1966), he measured formant values by using a ruler, paper and pencil. The x-ray contour tracings he showed are from a single speaker (himself) and were evaluated by visual inspection. The data collected for the current study are on the order of several GB per speaker (including, besides the ultrasound image data, synchronized audio and video of the lips). With the kind of procedures presented here the amount of work required in order to include additional speakers is potentially very small. Currently, the main manual task is to check the results of the automatic segmentation and alignment applied to the acoustic signal. Possibly even this stage could be skipped, if the strategy of recording as many speakers as possible is used to make the overall results robust to segmentation errors by sheer quantity of the data. It is thus hardly an overstatement to say that imaging techniques will be the vantage point for a new generation of speech production research: We can expect to see larger scale studies than ever before conducted even in small labs. In addition, the

quantification of articulatory movement during speech will supersede the nowadays typical focus on snapshots in time, bringing us towards a more comprehensive understanding of production dynamics.

Acknowledgments

Work supported by the ERC under the EU's 7th Framework Programme (FP/2007–2013) by Grant Agreement no. 283349-SCSPL and by Grant Agreement no. 295573-SCATS. Thanks to Jonathan Harrington and three anonymous reviewers for many constructive comments.

References

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

Barbosa, A.V., Déchaine, R.-M., Vatikiotis-Bateson, E., 2012. Quantifying time-varying coordination of multimodal speech signals using correlation map analysis. J. Acoust. Soc. Am. 131, 2162–2172.

Barbosa, A.V., Yehia, H., Vatikiotis-Bateson, E., 2008. Linguistically motivated movement behavior measured non-invasively. In: Proceedings of the International Conference on Auditory-Visual Speech Processing, AVSP, pp. 173–177.

Beautemps, D., Badin, P., Bailly, G., 2001. Linear degrees of freedom in production: analysis of cineradio- and labio-fim data and articulatory-acoustic modeling. J. Acoust. Soc. Am. 109, 2165–2180.

Berry, J., Fasel, I., 2011. Dynamics of tongue gestures extracted automatically from ultrasound. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 557–560.

Bladon, R.A.W., Al-Bamerni, A., 1976. Coarticulation resistance in English /l/. J. Phon. 4, 137–150.

Bombien, L., Hoole, P., 2013. Articulatory overlap as a function of voicing in French and German consonant clusters. J. Acoust. Soc. Am. 134, 539–550.

Bresch, E., Narayanan, S., 2009. Region segmentation in the frequency domain applied to upper airway real-time magnetic-image resonance images. IEEE Trans. Med. Imaging 28, 323–338.

Carignan, C., 2014. TRACTUS (Temporally Resolved Articulatory Configuration Tracking of UltraSound) Software Suite. http://christophercarignan.github.io/TRACTUS.

Carignan, C., Mielke, J., Dodsworth, R., 2016. Tongue trajectores in North American English /ae/ tensing. In: Côté, M.-H., Knooihuizen, R., Nerbonne, J. (Eds.), The Future of Dialects: Selected Papers from Methods in Dialectology XV (Language Variation 1). Language Science Press, Berlin.

Carignan, C., Shosted, R.K., Fu, M., Liang, Z.-P., Sutton, B.P., 2015. A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French. J. Phon. 50, 34–51.

Cederbaum, J., Pouplier, M., Hoole, P., Greven, S., 2016. Functional linear mixed models for irregularly or sparsely sampled data. Stat. Model. 16, 67–88.

Davidson, L., 2005. Adressing phonological questions with ultrasound. Clin. Linguist. Phon. 19, 619–633.

Davidson, L., 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. J. Acoust. Soc. Am. 120, 407–415.

Dawson, K.M., Tiede, M., Whalen, D., 2016. Methods for quantifying tongue shape and complexity using ultrasound imaging. Clin. Linguist. Phon. 30, 328–344.

Fasel, I., Berry, J.J., 2010. Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In: Proceedings of the 20th International Conference on Pattern Recognition, pp. 1493–1496.

Fu, M., Zhao, B., Carignan, C., Shosted, R.K., Perry, J.L., Kuehn, D.P., Liang, Z.-P., Sutton, B.P., 2015. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. Magn. Reson. Med. 73, 1820–1832.

Geumann, A., Kroos, C., Tillmann, H.G., 1999. Are there compensatory effects in natural speech? In: Proceedings of the 14th International Congress of Phonetic Sciences. San Francisco, USA, pp. 399–402.

Gubian, M., Torreira, F., Boves, L., 2015. Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. J. Phon. 49, 16–40.

Hall, K.C., Allen, C., McMullin, K., Letawsky, V., Turner, A., 2015. Measuring magnitude of tongue movement for vowel height and backness. In: Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow.

Harshman, R., Ladefoged, P., Goldstein, L., 1977. Factor analysis of tongue shapes. J. Acoust. Soc. Am. 63 (3), 693–707.

Harshman, R., Lundy, M., 1984. The PARAFAC model for threeway factor analysis and multidimensional scaling. In: Law, H.G., Snyder, C.W., Hattie, J.A., MacDonald, R.P. (Eds.), Research Methods of Mulitmode Data Analysis. Praeger, New York, pp. 122–215.

Hoole, P., 1999. On the lingual organization of the German vowel system. J. Acoust. Soc. Am. 106, 1020–1032.

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., 2007. Eigentongue feature extraction for an ultrasound-based silent speech interface. IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP07, Honolulu 1. pp. 1245–1248.

Iskarous, K., 2005a. Detecting the edge of the tongue: a tutorial. Clin. Linguist. Phon. 19, 555–565.

Iskarous, K., 2005b. Patterns of tongue movement. J. Phon. 33, 363–381.

Jacobs, M., Lehnert-LeHouillier, H., Bora, S., McAleavery, S., Dalecki, D., McDonough, J., 2008. Speckle tracking for the recovery of displacement and velocity information from sequences of ultrasound images of the tongue. In: Proceedings of the 8th International Seminar on Speech Production, Strasbourg pp. 53–56.

Kisler, T., Reichel, U., Schiel, F., Draxler, C., Jackl, B., Pörner, N., 2016. BAS speech science web services. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. paper id 668.

Ladefoged, P., Maddieson, I., 1996. The Sounds of the World's Languages. Oxford Blackwell.

Lammert, A., Proctor, M., Narayanan, S., 2010. Data-driven analysis of realtime vocal tract MRI using correlated image regions. In: Proceedings of Interspeech, pp. 1572–1575.

Lancia, L., Rausch, P., Morris, J.S., 2015. Automatic quantitative analysis of ultrasound tongue contours via wavelet-based functional mixed models. J. Acoust. Soc. Am. 137, EL178.

Lancia, L., Tiede, M., 2012. A survey of methods for the analysis of the temporal evolution of speech articulator trajectories. In: Fuchs, S., Weirich, M., Pape, D., Perrier, P. (Eds.), Speech Planning and Dynamics. Peter Lang, Frankfurt, pp. 239–277.

Li, M., Kambhamettu, C., Stone, M., 2005. Automatic contour tracking in ultrasound images. Int. J. Clin. Linguist. Phon. 19, 545–554.

Lin, S., Beddor, P.S., Coetzee, A.W., 2014. Gestural reduction, lexical frequency, and sound change: a study of postvocalic /l/. Lab. Phonol. 5, 9–36.

Maeda, S., 1990. Compensatory articulation during speech; evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Dordrecht, Kluwer, pp. 131–150.

McMillan, C., Corley, M., 2010. Cascading influences on the production of speech: evidence from articulation. Cognition 117, 243–260.

Ménard, L., Aubin, J., Thibeault, M., Richard, G., 2012. Measuring tongue shapes and positions with ultrasound imaging: a validation experiment using an articulatory model. Folia Phoniatr. Logop. 64, 64–72.

Mielke, J., 2015. An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons. J. Acoust. Soc. Am. 137, 2858–2869.

Moisik, S., Lin, H., Esling, J., 2014. A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). J. Int. Phon. Assoc. 44, 21–58.

Morris, J.S., Carol, R.J., 2006. Wavelet based functional mixed models. J. R. Stat. Soc. Ser. B 68, 179–199.

Mücke, D., Grice, M., Cho, T., 2014. More than a magic moment – paving the way for dynamics of articulation and prosodic structure. J. Phon. 44, 1–7.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D., 2004. An approach to real-time magnetic resonance imaging for speech production. J. Acoust. Soc. Am. 115, 1771–1776.

Niebergall, A., Zhang, S., Kunay, E., Keydana, G., Job, M., Uecker, M., Frahm, J., 2012. Real-time MRI of speaking at a resolution of 33 ms: undersampled radial FLASH with nonlinear inverse reconstruction. Magn. Reson. Med. 69, 477–485.

Öhman, S.E., 1966. Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Am. 39, 151–168.

Parell, B., Narayanan, S., 2014. Interaction between general prosodic factors and language-specific articulatory patterns underlies divergent outcomes of coronal stop reduction. In: Proceedings of the 10th International Seminar on Speech Production, Cologne, Germany, pp. 308–311. 5–8 May 2014.

Pouplier, M., 2008. The role of a coda consonant as error trigger in repetition tasks. J. Phon. 36, 114–140.

Proctor, M., 2011. Towards a gestural characterization of liquids: evidence from Spanish and Russian. J. Lab. Phonol. 2, 451–485.

Proctor, M., Bone, D., Katsamanis, A., Narayanan, S., 2010. Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In: Proceedings of *Interspeech*. Makuhari, Japan, pp. 1576–1579.

Proctor, M., Goldstein, L., Lammert, A., Byrd, D., Toutios, A., Narayanan, S., 2013. Velic coordination in French nasals: a real-time magnetic-resonance imaging study. In: Proceedings of *Interspeech*.

Proctor, M., Lammert, A., Katsamanis, A., Goldstein, L., Hagedorn, C., Narayanan, S., 2011. Direct estimation of articulatory kinematics from real-time magnetic resonance image sequences. In: Proceedings of Interspeech, Florence, pp. 281–284.

Recasens, D., 2012. A cross-language acoustic study of initial and final allophones of /l/. Speech Commun. 54, 368–383.

Recasens, D., Espinosa, A., 2009. An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. J. Acoust. Soc. Am. 125, 2288–2298.

Recasens, D., Pallarès, M.D., Fontdevila, J., 1997. A model of lingual coarticulation based on articulatory constraints. J. Acoust. Soc. Am. 102, 544–561.

Schiel, F., 1999. Automatic phonetic transcription of non-prompted speech. In: Proceedings of International Congress of Phonetic Sciences, ICPhS, San Francisco, August 1999, pp. 607–610.

Scobbie, J., Wrench, A., van der Linden, M., 2008. Head-probe stabilization in ultrsound tongue imaging using a headset to permit natural head movement. In: Proceedings of the 8th International Seminar on Speech Production, Strasbourg, pp. 373–376.

Silva, S., Teixeira, A., 2015. Unsupervised segmentation of the vocal tract from real-time MRI sequences. Comput. Speech Lang. 33, 25–46.

Smith, C., 2014. Complex tongue shaping in lateral liquid production without constriction based goals. In: Proceedings of the 10th International Seminar on Speech Production, Cologne, Germany, 5–8 May 2014, pp. 413–416.

Stone, M., 2005. A guide to analysing tongue motion from Ultrasound images. Int. J. Clin. Linguist. Phon. 19, 455–501.

Stone, M., Faber, A., Raphael, L.J., Shawker, T., 1992. Cross-sectional tongue shape and linguopalatal contact patterns in [s], [sh], and [l]. J. Phon. 20, 253–270.

Toda, M., 2009. Etude Articulatoire et Acoustique des Fricatives Sibilantes. Ph.D. Thesis Université Paris.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. J. Cogn. Neurosci. 3, 71–86.

Uecker, M., Zhang, S., Voit, D., Karaus, A., Merboldt, K.-D., Frahm, J., 2010. Real-time MRI at a resolution of 20 ms. NMR Biomed. 23, 986–994.

Wieling, M., Montemagni, S., Nerbonne, J., Baayen, R.H., 2014. Lexical differences between Tuscan dialects and standard Italian: accounting for geographic and sociodemographic variation using generalized additive mixed modeling. Language 90, 669–692.
Wrench, A., 2008. Articulate Assistant Advanced User Guide v.2.05. www.articulateinstruments.com.
Wrench, A., Scobbie, J., 2006. Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. In: Proceedings of the 7th International Seminar on Speech Production. Ubatuba, Brazil, pp. 451–458.
Yip, J., 2013. Phonetic Effects on the Timing of Gestural Coordination of Modern Greek Consonant Clusters. University of Michigan, Ann Arbor.