Investigating articulatory variation in the DoubleTalk corpus: Electromagnetic Articulography evidence of phonetic and speech style variation

B037795

8,015 words

MSc Speech and Language Processing
University of Edinburgh

2017

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*B037795*)

For my family

Acknowledgements

With thanks to my supervisor Dr. James Kirby, and additional input from Dr. Korin Richmond, for their guidance during the composition of this project. Thanks also for the support from my friends in the DSB lab: Akela Drissner, Kai Konkolewski (diolch yn fawr), Jonas Robertson, Maureen de Seyssel, Jason Taylor, and Simon Vandieken.

Abstract

The present investigation expands on differences in speech articulation between read and spontaneous speech tasks in the DoubleTalk corpus. This corpus comprises of electromagnetic articulography data from multiple participants in a high temporal resolution. I analysed speech style differences on the phone level, using a subset of phones, speech tasks, and research participants. I observed qualitative differences in the articulatory space for each phone and each individual between speech styles, and performed a quantitative analysis on one participant's data. These observations are discussed in the context of theories of articulation, statistical identification of articulators, and other electromagnetic articulography investigations.

Keywords: Articulation, EMA, speech, variation

Contents

Chapter 1

Introduction

Instrumental techniques in phonetics such as electromagnetic articulography (EMA) allow more fidelity in collecting physiologically real speech data. It also captures articulation which may remain undetected in a raw acoustic analysis, and has shed light on the nature of co-articulation (Harrington, Fletcher, & Roberts, 1995), assimilation (Ellis & Hardcastle, 2002), and prosodic (Cho, 2006) and dialectal differences in articulation (Wieling et al., 2016). The data from instrumental techniques provides us with concrete information to supplement abstract ideas of human speech production and perception. It is also useful for speech technology systems like Automatic Speech Recognition (King et al., 2007), and features of the speech signal which are useful in synthesising speech.

This investigation expands on findings from Parsons' investigation into acoustic-to-articulatory inversion (2015), where she found differences in articulatory trajectories between read and spontaneously-produced speech. She found that spontaneous speech was less precise in the articulatory space, and the articulators moved towards their targets at a slower velocity. This study incorporates the same research participants from the DoubleTalk corpus (Scobbie et al., 2013), but analyses a subset of phones for each speaker to investigate differences between speech style for each phone, as well as providing accompanying quantitative analyses.

This study begins with an overview of instrumental phonetic methods, as well as an in-depth description of EMA. It then introduces aspects of phonological theory related to speech planning, coarticulation, and its impact on speech style, including an overview of critical articulation, a statistical approach. The following section describes my data collection and analysis, and then presents qualitative and quantitative articulatory

data. The final section addresses the implication of my findings in relation to previous data and articulatory theories. It also proposes further areas of research brought about by this investigation.

Chapter 2

Background

2.1 Capturing articulation

A trade-off exists in methods of capturing articulation - imaging of the vocal tract and point-tracking. Methods of imaging include Ultrasound Tongue Imaging (UTI) and Magnetic Resonance Imaging (MRI). These provide a more faithful image of the vocal tract, but at a lower temporal resolution. UTI and MRI typically capture images at 30-120fps[1] and 50-100fps respectively (Lawson et al., 2015). UTI has the benefit of both being relatively inexpensive, easy to set up, and non-invasive (Lawson et al., 2015). However, images of the tongue are often unclear, and ultrasound is unable to reflect of the hard tissue of the palate and the teeth - as well as not capturing articulation of the lips (Gick, Wilson, & Derrick, 2012, p. 160). MRI produces faithful images of the entire vocal tract except the teeth (Gick et al., 2012, p. 223). However, MRI is highly expensive and involves participants speaking in confined, unnatural spaces where the tongue's resting position is altered.

Point-tracking methods produce representations of the articulators at higher resolutions. One method is to use infrared-emitting diodes such as in the *Optotrak* system. These are attached to nine locations on the face and capture movement at 460fps (Gick et al., 2012, p. 200). However the tongue, an articulator crucial for most speech sounds, is not measured due to the non-invasive nature of *Optotrak*. Another method which offered both high temporal and spatial resolution was x-ray microbeam. This methods involved placing a string of metal (usually gold or lead-based) spheres along the mid-sagittal plane of the tongue and firing low levels of concentrated x-rays through

[1] frames per second

participants' heads (Westbury, Turner, & Dembowski, 1994). This system's apparatus is large and highly expensive to operate (Gick et al., 2012, p. 201), is no longer in operation, and is unlikely to pass modern ethics criteria due to the ionising radiation that participants are exposed to. In addition, data visualisation and quantification is only available in 2D (Papcun et al., 1992).

2.1.1 Electromagnetic articulography

Since the discontinuation of x-ray microbeam equipment, EMA systems have been used to construct multiple corpora. EMA involves gluing up to twelve sensor coils to speech articulators and reference points on a speaker's head. These sensors are detected by six transmitter coils which generate alternating magnetic fields. The transmitter coils produce carrier frequencies of between 7.5-13.5kHz (Geng et al., 2013). Like all other methods mentioned above to capture articulation, the acoustic signal may also be picked up with an accompanying microphone. Due to the electromagnetic field in which the participant is sat, it is necessary to use a piezoelectronic microphone which contains no moving metal coils. This is due to the electromagnetic interference a conventional microphone would produce (Geng et al., 2013).

Figure 2.1 shows the typical placement of the sensor coils, as is the case in the DoubleTalk corpus (Scobbie et al., 2013). Coils for capturing speech articulation are Upper Lip (UL), Lower Lip (LL), Lower Jaw central (LJ), Tongue Tip (TT), Tongue Body (TB), and Tongue Dorsum (TD) along the mid-sagittal plane. Due to non-exact placement and the physiology of every individual as unique, sensor coils are treated in relative position to each other, and other coils used for reference positions. These reference coils, indicated with 'ref' in Figure 2.1, are located behind each ear, on the upper incisors, and on the bridge of the nose. These coils are important for system calibration and as a reference point in EMA data analysis.

While older EMA systems could only measure movement in three-dimensional space (x,y,z) (Perkell et al., 1992), newer systems - such as those used to create the MOCHA (Wrench, 2000), DoubleTalk, and mngu0 (Steiner, Richmond, Marshall, & Gray, 2012) corpora - can measure five degrees of freedom (Gick et al., 2012, p. 144). These include movement in a three-dimensional plane (x,y,z), and two degrees of rotation to more accurately reflect the motion of the tongue. EMA trades off its high temporal resolution of 200fps and its five degrees of freedom with the imaging of up to only

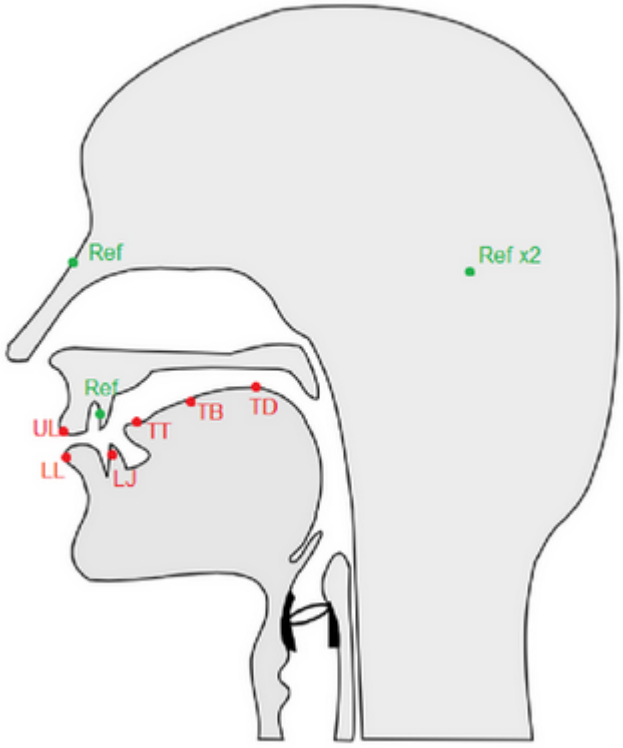


Figure 2.1: Typical EMA sensor coil placement locations. Red = articulators for measurement, green = reference points

twelve points. This means high temporal and spatial accuracy, but with a higher level of visual abstraction than methods like UTI and x-ray microbeam.

2.2 Analysing articulation

Prior to the development of accurate instrumental techniques, articulation was inferred from the acoustic signal, or articulatory theories were developed from models and images of the vocal tract. This section focuses on how theories of articulation may account for differences in read versus spontaneously produced speech. These are based on varying approaches to the phenomenon of coarticulation, and experimental findings that spontaneous speech is globally different in acoustic (Nakamura, Iwano, & Furui, 2008) and articulatory (Parsons, 2015) character to read speech. Moreover, spontaneous speech results in lower reconigtion rates in Automatic Speech Recognition (ASR) systems (Nakamura et al., 2008).

Coarticulation is described as the overlap of speech articulators in time, where the configuration of the vocal tract during the production of a phone is influenced by adjacent phones. In addition, these configurations often undertake similar articulatory strategies, and may be below the level of acoustic or human perception (Farnetani & Recasens, 2010).

2.2.1 Context-sensitivity

It has been widely accepted that modifying stress patterns, speech rate, and ideas of 'attention paid to speech' (e.g., Coupland, 1980), affect both the acoustic signal (Tuller, Harris, & Kelso, 1982; Matthies, Perrier, Perkell, & Zandipour, 2001) and therefore the configuration of the speech articulators. In particular, speech patterns involving pausing and breathing significantly vary between planned and unplanned utterances (Grosjean & Collins, 1979). This has implications in read versus spontaneous speech, where the articulatory goals rely on both temporal constraints, and the degree to which the utterance has been planned.

Within the literature, researchers disagree about how the articulators behave in reaching their targets. Targets or goals in this instance are defined as the articulatory configuration required to produce and perceive a given phone. Context-sensitive approaches require movements to be planned according to the following phones in a sequence, whereas context-insensitive approaches rely on the motor system to average

out trajectories into articulatory configurations where coarticulation occurs (Gick et al., 2012, p. 207). There are also hybrid approaches, combining ideas of context-sensitivity.

The Task Dynamics model of Articulatory Phonology (AP/TD) (Browman & Goldstein, 1992) is considered a context-insensitive model (Gick et al., 2012, p. 208). Articulatory gestures are seen as reliant on 'constellations' of articulators (for example, larynx, tongue tip, velum *etc.*), where not all articulators are required to produce a given phone. Because these constriction degrees and locations are goals, temporal constraints are considered intrinsic to the model. This means that artiulatory patterns may overlap over time, but this means that instead of accurately mapping the articulators, there is still a level of abstraction required to understand the model. Another hybrid approach, the DIVA model (e.g., Perkell et al., 2000), is a response to AP/TD. Perkell and other DIVA researchers argue that AP/TD only accounts for speech production, while there must be consideration for the hearer. Temporal constraints in speech are also intrinsic to the DIVA model, and speech planning takes into account the physiology of the vocal motor system. For example, rate of speech may lead to 'inertia' (Perkell et al., 2000) - actual articulatory trajectories lagging behind planned ones. This means that the vocal motor system must use 'feedforward' articulatory planning (Perkell, 2012) to produce a configuration perceivable as the target phone to the hearer (c.f., Liberman & Mattingly, 1985).

Variable configurations due to physiological and temporal constraints have implications to how articulators behave between speech styles. Read speech, which is more likely to be carefully produced and pre-planned, should - according to the theories outlined above - be articulated differently to spontaneous speech.

2.3 Critical Articulation

Once articulatory data from sources like EMA were available to researchers, they could employ statistical methods of studying patterns of articulation. Studying critical articualtors involves identifying which articulator configurations are considered 'critical', 'dependent' or 'redundant' in producing a given phone (Singampalli & Jackson, 2007). Singampalli and Jackson (2007) developed an algorithm to identify, as well as measure how strong the influence of, articulators in producing target phones. Their algorithm was based on both a univariate and bivariate model. The univariate model analysed 1-dimensional correlations between the trajectories of the upper and lower lips, tongue

tip, body and dorsum, the lower jaw, and the velum. These were the EMA sensor coils from the MOCHA-TIMIT corpus (Wrench, 2000), the training data for Singampalli and Jackson's algorithm. The bivariate model used all of the same articulators, but two extra degrees of freedom - the x^2 and y axes - were taken into account. The identification of critical and dependent articulations is based on a distance measure between articulatory coordinates (Jackson & Singampalli, 2009), Kullback-Leibler divergence. Critical articulator studies have identified articulators which are crucial to the production of a range of IPA phone categories and differ slightly between male and female participants (Jackson & Singampalli, 2009). In addition, they have investigated articulator movement trajectories (Kim, Lee, & Narayanan, 2014), differences in emotion in speech (Kim, Toutios, Lee, & Narayanan, 2015), and the utility of critical articulation to ASR acoustic features (Felps, Geng, Berger, Richmond, & Gutierrez-Osuna, 2010).

It will be important to reference critical articulator studies in this investigation. They provide guidelines into which articulators to pay close attention to during analysis of the read and spontaneous EMA data.

2.4 EMA studies

EMA studies have uncovered articulatory differences in speech which are difficult or impossible to identify in acoustic or auditory analyses. For example, Cho (2006) uncovered differences in lip trajectories, where articulations at prosodic boundaries displayed less overlapping and a greater degree of opening. His analysis was supplemented with concepts of gestural overlap from AP/TD theories (Browman & Goldstein, 1992). Isakrous and colleagues (2011) also found that phrase and word position, from variation in /s/, was responsible for variation in lower jaw configuration - as well as a trade off in the movement trajectories of the tongue tip and tongue dorsum. The implications for the present investigation is that in spontaneous speech, where there is likely to be a greater level of articulatory overlap, phones are likely to be produced with a smaller degree of lip-opening.

Along with speech style differences, dialect variation has also been shown to result in predictably variable articulatory configuration. Wieling and colleagues (2016) noticed that there was a north/south divide in Dutch, where the tongue's position was considered more anterior for southern speakers across all of their read data. Moreover,

[2] becomes z axis in 3D plots

the critical articulator study discussed in section 2.3 observed differences in articulator configuration for certain IPA phone categories between genders, male and female, in the MOCHA-TIMIT corpus (Jackson & Singampalli, 2009). These studies show that differences in speech style, gender, and dialect all have an impact on articulation. Therefore, investigating differences between different DoubleTalk speech tasks should result in variation between read and spontaneous exercises.

2.4.1 Parsons study

The motivation of this investigation is based on an epiphenomenal finding in Parsons' (Parsons, 2015) thesis on acoustic-to-articulatory inversion. Parsons' study also used the DoubleTalk corpus and found that, between read and spontaneous speech, there were visibly different articualtor configurations. An example plot from her study is shown in Figure 2.2. This plot shows the articulators from a 'front-on' perspective, focusing on the x-axis where the articulators are less variable than on the other axes. In this investigation, another viewing angle is used. Across all data, including non-speech pauses, there appeared to be a greater degree of variation in lip and tongue movement. Parsons observes that the articulators move with greater velocity and in 'smaller' gestural movements, while spontaneous speech appeared more imprecise; it uses a greater area of the articulatory space and with slower gestural velocity (Parsons, 2015). The nature of Parsons' study did not allow closer investigation, so it is valuable to undertake a closer investigation of individual phones for each speaker to analyse the differences between read and spontaneous articualtion. In addition, as Parsons modelled the entirety of all speech files, variation in articulator movement might be due to non-speech cues.

2.5 Articulatory phonetics and speech technology

The field of speech technology is only beginning to adopt the current understanding of human speech production. Even though articulatory data is becoming more abundant "and may be regarded as ground truth, it is not sufficient to build a *model* (King et al., 2007, p.5)". Acoustic models in systems such as Hidden Markov Model (HMM) recognisers rely on linear sequences of phones which are based on a phonetic alphabet (Jurafsky & Martin, 2014). The problem with this approach is that many traces in an acoustic signal can produce the same phone category in a language, and broad phonetic

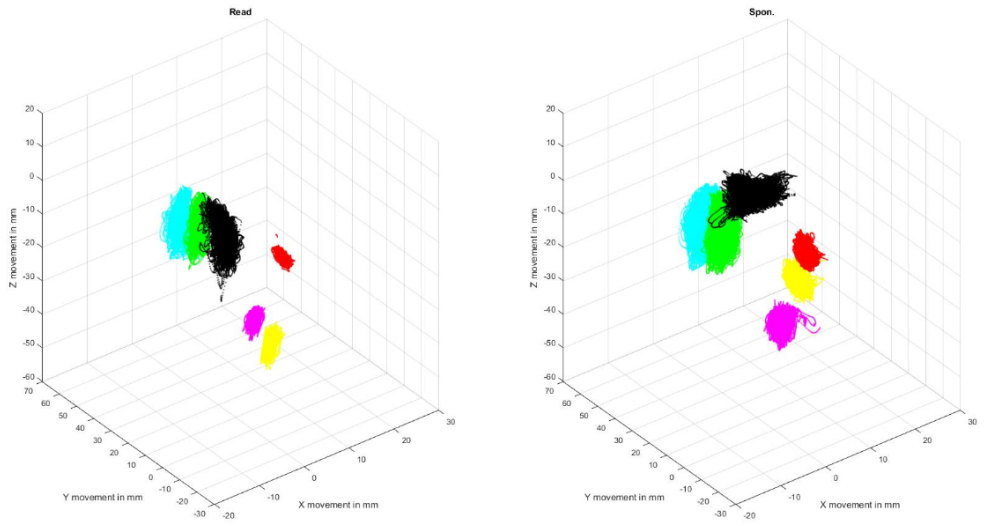


Figure 2.2: Example plot from Parsons (2015). The articulators are all coloured differently, and use the same key as the scatterplots in section 4.1. All articulator coils pictured are located on the midsagittal plane

categories may be configured differently across languages. Acoustic-to-articulatory inversion (c.f. Papcun et al., 1992; Parsons, 2015) is one approach to inferring underlying articulatory configuration from the acoustic signal. Articulatory patterns for individual features and phones are mathematically correlated from aspects of the speech signal using statistical methods such as HMMs, or by using the hidden layers of neural networks. By calcualting root mean square error, Parsons (2015) discovered that read speech more faithfully revealed underlying articulation than spontaneous speech. A closer investigation into individual phones allows a closer investigation into how her neural network used the acoustic data to map articulator configurations, and the reasons why read speech is more useful to the inversion process.

2.6 Research questions

Motivated by existing EMA studies, the study of critical articulators, and theories relating to the effects of coarticulation, my investigation is based around the following research questions:

- Is there articulatory variation between read and spontaneous speech on the phone level?

- How is articulatory variation between phones manifested?

Chapter 3

Methodology

In order to investigate differences between read and spontaneous speech, I gathered EMA data from a subset of the DoubleTalk corpus. The entire corpus was not used so that the duration of read and spontaneous speech for each participant was roughly equal, and due to time constraints in data pre-processing. Out of the twelve speakers comprising the DoubleTalk corpus, I used data from six participants due to a range of text-labelling alignment and sensor coil errors (Parsons, 2015).

3.1 The DoubleTalk corpus

The DoubleTalk corpus (*full description* - (Scobbie et al., 2013)) was collected at the Edinburgh Speech Production Facility between 2008-2010. This facility is unique in that two synchronised Carstens AG500 electromagnetic articulometers are set up so that speech can be recorded in dialogue. The two machines are placed far enough apart to cause as little electromagnetic interference as possible to each other, and are equipped with talkback piezoelectronic microphones (Geng et al., 2013) at a sampling rate of 32kHz (Parsons, 2015). Every participant had the ten sensor coils shown in Figure 2.1 glued to their articulators with dental adhesive and coated in latex to prevent them from becoming de-attached during recording. Ten of the twelve speakers had an additional 'lower jaw lateral' sensor placed on the bottom of their chin (Parsons, 2015) in order to measure the mouth's degree of opening.

The corpus included speech tasks in both monologue and dialogue. The first task was a reading of the *Comma Gets a Cure* (McCullough, Somerville, & Honorof, 2000) passage modified for the extended lexical sets of Scottish Standard English (Wells,

Table 3.1: DoubleTalk participants used for analysis. GN = General Northern (anglo-English), RP = Received Pronunciation, SSE = Scottish Standard English, SSBE = Standard Southern British English

Speaker	Gender	Dialect	Read tokens	Spontaneous tokens
r20cs5	Male	GN	496	707
r33cs6	Male	SSBE	484	1,008
r34cs6	Female	RP	466	493
r35cs5	Male	SSE	493	1,055
r35cs6	Male	SSBE	490	632
r36cs5	Female	SSE	504	876

1982), as well as a word list. Participants were also asked to perform a spontaneous monologue drawing on an anecdote from their own life. The interactive tasks in dialogue involved *spot the difference* and map task games, as well as a recall task where the other member of a dyad was asked to retell their interlocutor's anecdote (Scobbie et al., 2013).

The varied and large amount of data in dyads is currently unique to EMA corpora, and is especially advantageous to this research project. In the following sections, I describe the subset of data I analysed, the method of data processing I undertook, and the motivations for my use of this data.

3.1.1 Participants

From the remaining speakers where there were no errorful *.TextGrid* and *.pos* sensor files, there remained a good balance in gender and dialect distribution. Speaker information and number of tokens gathered from each is displayed in Table 3.1. All of the DoubleTalk speakers apart from one pair are reported as naive non-linguists who did not know each other prior to recording (Scobbie et al., 2013). One speaker in my subset (r20cs5) is one of the linguist, non-naive participants. All participants did not report speech or hearing deficits.

3.1.2 Materials

My subset of DoubleTalk comprises of one individual, read task - the *Comma Gets a Cure* passage, and one interactive, spontaneous task - *spot the difference*. For all six participants, the total read data duration was 18m18s with an average duration of 3m03s between participants. The total duration for spontaneous data was 45m06s, with an average duration of 7m31s between participants. The amount of tokens per task, per speaker is displayed in Table 3.1.

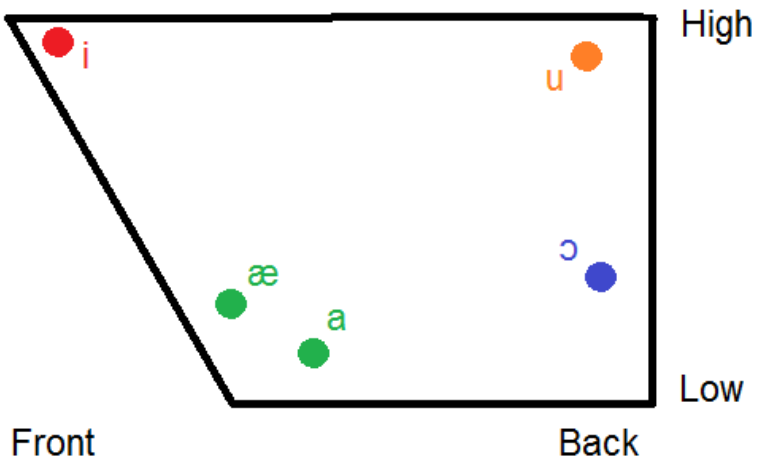


Figure 3.1: Vowels used for articulatory speech style analysis within the vowel space. An abstract representation based on the IPA

To analyse differences between speech styles, I measured four vowels and three plosives for each speaker. I attempted to have a wide range in place of articulation for plosives by choosing /p,b/, /t,d/, and /k,g/. This is also an attempt to analyse the variation in particular articulators such as the upper and lower lips for /p,b/, the tongue tip for /t,d/ and the tongue dorsum for /k,g/. As for vowels, I measured /i/, /a,æ/, /ɔ/, and /u/. These vowels were chosen in order to parsimoniously represent tongue position for the entire English vowel space (e.g., Lindblom, 1986), and is shown pictorially in Figure 3.1. /a/ and /æ/ are measured as the same variable due to the ARPAbet representation for both phones being the same, /AE/.

3.2 Data pre-processing

The TextGrid files accompanying DoubleTalk only label each sound file with a phrasal transcription (Scobbie et al., 2013). It was therefore necessary to generate a phone-level transcription for my subset of DoubleTalk data. This was achieved by running the FAVE-aligner (Rosenfelder, Fruehwald, Evanini, & Yuan, 2011), which uses a Hidden Markov Model method of forced alignment. The software requires a specifically-formatted *.txt* file and accompanying *.wav* file, and returns a *.TextGrid* file with a word-level and phone-level transcription. The phones are labelled with ARPAbet transcription from entries in the Carnegie Mellon University dictionary. Within the FAVE software suite, a Praat (Boersma & Weenink, 2005) script, *Convert_To_FAVE-*

align_Input.praat, is used to produce the *.txt* file readable by the FAVE-aligner. One technical issue encountered with the FAVE-aligner is that, though words and speech errors - such as false starts - without a dictionary entry may be added manually to a transcription, silences cannot. Silences already encoded between phrases in the original transcriptions are therefore present, but silences between words and within phrases are not. Overall, the transcriptions produced by the FAVE-aligner were accurate across read and spontaneous files.

In order to plot the transcribed phones from the EMA *.pos* data, I needed to align the frame representations from the *.pos* file with the TextGrid timetamps. First, this involved manipulating the 3D binary arrays in which the articulatory data is encoded. This was achieved by creating *numpy* arrays which contain each 5ms frame, 10-12 EMA sensor coil channels, and five coordinates which map the position and rotation of each coil. The code necessary to create these 3D matrices is included in my *posplot.py* script in Appendix A. One issue with the 3D matrices is that the EMA channel for each sensor varies between speakers. I consulted Parsons (2015, p. 48) to match the appropriate sensor with each articulator for each participant.

Next, it was necessary to extract the desired phones for analysis from the TextGrid files, and convert the second value of their start and end-points into the frame number within each *.pos* file. In the *framewav.py* script (Appendix B), I used the PraatIO[1] module to extract all instances of a given phone with their start and endpoint timestamps. I then converted their second values into frame numbers within their *.pos* files. This was achieved by calculating Equations 3.1 and 3.2

$$f_x = \frac{s_{secs}}{T_{secs}} \cdot F \tag{3.1}$$

$$f_y = \frac{e_{secs}}{T_{secs}} \cdot F \tag{3.2}$$

Where f_x = frame ID for a given token's startpoint, f_y = frame ID for a given token's endpoint, s_{secs} = startpoint in seconds, e_{secs} = endpoint in seconds, T_{secs} = total filelength in seconds, and F = total filelength in frames.

[1] https://github.com/timmahrt/praatIO

3.3 Data analysis

Once I obtained all of the frame IDs for the desired phone classes, I plotted each phone for each speaker in 3D scatter plots. These plots are in a similar format to Parsons (2015), the motivation for this investigation, but are viewed at from a different angle. This is because there is little movement on the x-axis due to the physiology of the oral cavity, and the fact that all sensors are located midsagittally.

Movement along the x axis is lateral, left-to-right, movement. The y axis depicts front-back movement, and the z axis depicts longitudinal, top-to-bottom, movement. The axes in each plot are equal in scaling, and movement is measured in millimetres.

As displayed and discussed in Parsons (2015), I qualitatively analysed differences between articulatory movements in read and spontaneous speech. I expand on this work by discussing differences between phones. I then quantitavely analysed the distribution of articulator positions at phone midpoints for one speaker. I chose to analyse participant r35cs6 for concision, and also because his data was free from NaN[2] errors in the *.pos* files.

I performed paired t-tests to consider whether articulator distributions significantly differ between speech styles. Where there is a significant difference between the distribution of articulator configuration, it can be assumed that speech style has an affect on speech production.

3.4 Reproducibility

As of August 2017, DoubleTalk materials are available online at http://espf.ppls.ed.ac.uk/. The FAVE software suite is also available via GitHub (https://github.com/JoFrhwld/FAVE). The scripts I wrote to process and plot DoubleTalk data are in Appendices A and B.

As all of the materials are freely available, there is a scope for other researchers to plot a different selection of phones and speech tasks for speakers within the DoubleTalk corpus. Moreover, my data has contributed phone transcriptions to a subsection of the corpus and will be presently included online.

[2] 'not a number'

3.5 Predictions

From Parsons' findings (2015), studies of critical articulation (Singampalli & Jackson, 2007; Kim et al., 2014) and previous work on dialectal (Wieling et al., 2016) and prosodic (Cho, 2006) variation, I expect there to be clear differences in the articulation of read and spontaneous speech. An interesting line of enquiry is connecting differences in speech style and critical articulation. I predict that many articulators will be both qualitatively and statistically variable in the production of each phone for each speech style, especially with the articulators considered critical. The articulators considered critical for the phones I investigated are located in Table 3.2.

Table 3.2: Articulators and direction of movement considered critical (Jackson & Singampalli, 2009) for the phones I analysed. -Y and -Z refer to the axis of articulator movement

Phone	'Critical' for articulation
/p,b/	UL-Y, LL-Y
/t,d/	TT-Y, LJ-Y(/t/, female only)
/k,g/	TD-Y
/i/	TT-Z
/a,æ/	LL-Y
/ɔ/	TB-Y, TB-Z, TD-Z
/u/	none

Chapter 4

Results

The main finding of this investigation is that there is more supporting evidence for articulatory differences between read and spontaneous speech. First, I present the 3D representations of articulatory space for each speaker and make qualitative comparisons on the distribution of the data. Then, I present the results of my t-tests on each articulator for each phone for speaker r35cs6. I compare these findings with my predictions and previous critical articulator studies. I finally present articulator trajectory data from Parsons (2015).

4.1 Raw articulatory data

Figures 4.1-4.6 show scatter plots of the articulatory trajectories for the seven phones investigated for each participant. The articulators are colour-coded and are included in each plot's figure legend. The nose is included as a point of reference in all plots. Like Parsons' comparison between read and spontaneous data (2015), there appears to be a greater degree of front-back (y-axis) and longitudonal (z-axis) articulatory movement, particularly in the tongue for spontaneous speech. However, Parsons' interpretation that the lips are closer together in spontaneous speech is only apparent in certain speakers and certain phones. A clear example of this is in speaker r36cs5 in Figure 4.6. For all phones, particularly /ɔ/, there is less space on the longitudonal plane between the UL and LL sensor traces. Parsons' other observation that the articulators are more 'precise' in read speech is especially shown in low-frequency tokens, particularly /u/. Clear examples are shown when comparing the tongue sensors between read and spontaneous /u/ for r20cs5, r33cs6, and r34cs6, and r35cs6. There is much more

variation on the z-axis for all of these speakers' realisations of /u/. Fewer tokens may have constrained the preceding and following phone environments which would, in turn, constrain articulatory configuration futher.

Along the y-axis, the variation in front-back tongue movement is demonstrated by the overlap in tongue articulator trajectories. This is particularly clear in participants r20cs5, r33cs6, and r34cs6. A greater degree of longitudonal variation is most clear in participants r33cs6, r34cs6, r35cs5, and r36cs5. As in Parsons (2015), longitudonal variation is clearly manifested in the tongue tip, but there are also clear differences in the lower lip, lower jaw, and other tongue articulators.

These plots also show that there is interspeaker variation in the configuration of the articulators across all phones, particularly the tongue. These differences are manifested in the tongue tip and seem to remain constant across both speech styles, but vary between participants. The tongue tip is generally lower than the tongue body and tongue dorsum for participants r20cs5, r35cs5, and r36cs5. For participants r34cs6 and r35cs6, the tongue tip remains level with the other tongue sensors, and for participant r33cs6, the tongue tip is generally raised above the other tongue sensors.

An additional finding between read and spontaneous speech is from lateral movement for participants r34cs6 and r36cs5. In their spontaneous data, there are a few articulator trajectories which appear to be away from the midsagittal plane on the x-axis. This may infer articulations where the participant's entire head shifts to the left or right during the production of spontaneous speech, or that the *.pos* file has misinterpreted positioning coordiantes (Richmond, *p.c.*).

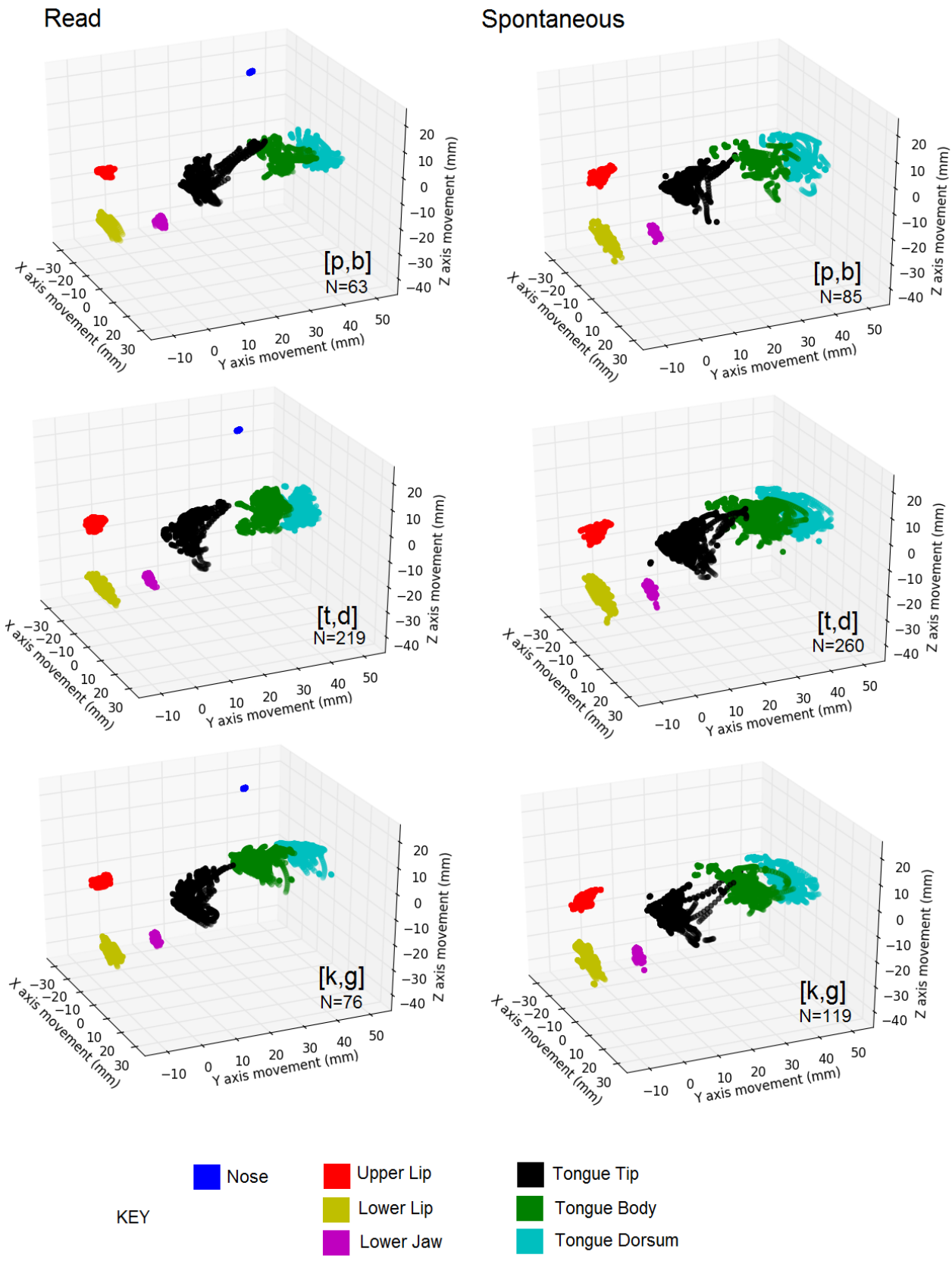
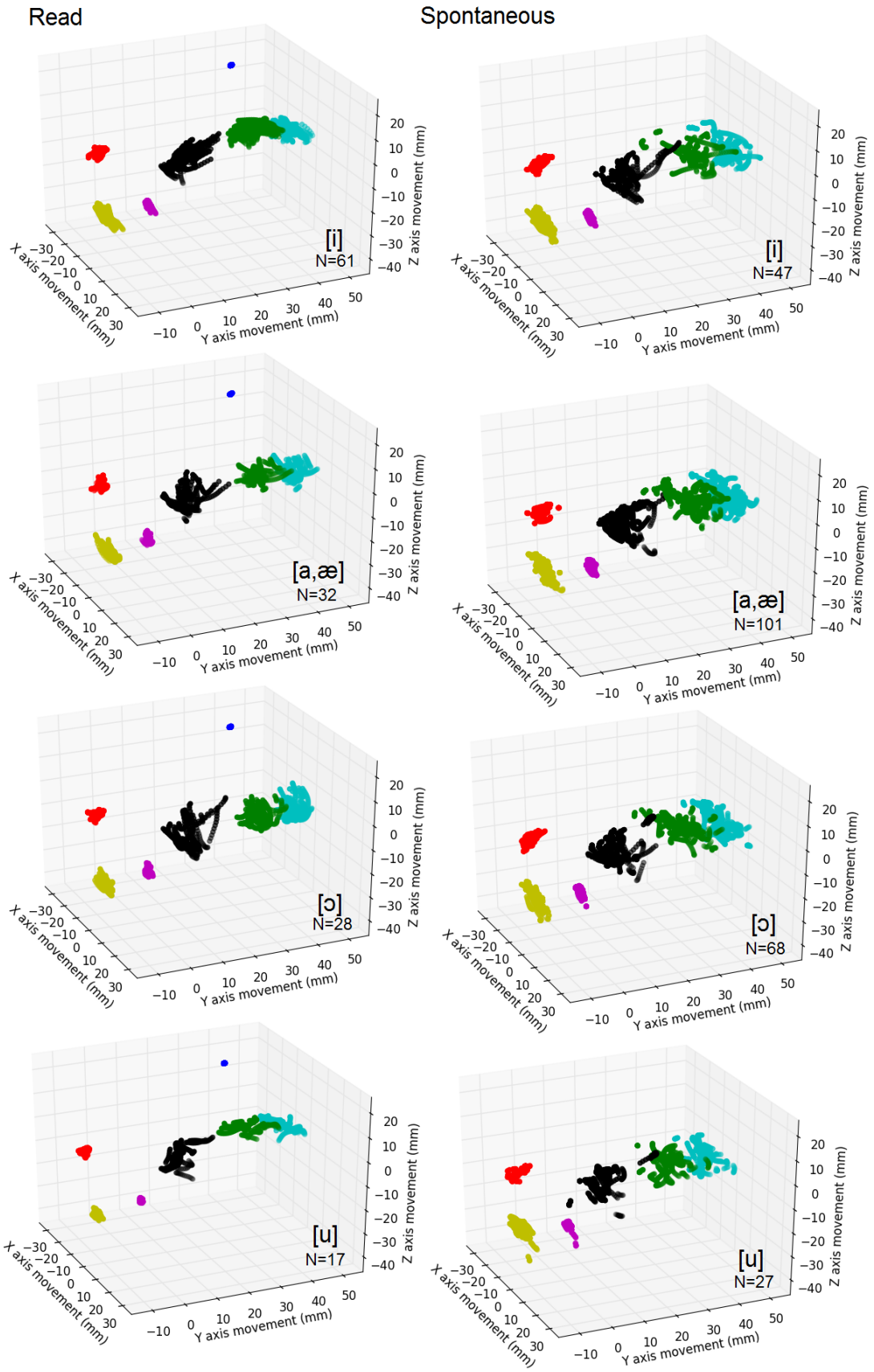


Figure 4.1: Articulator trajectories for speaker r20cs5 - General Northern, male


Read
Spontaneous
[i]
N=61
[i]
N=47
[a,æ]
N=32
[a,æ]
N=101
[ɔ]
N=28
[ɔ]
N=68
[u]
N=17
[u]
N=27
X axis movement (mm)
Y axis movement (mm)
Z axis movement (mm)

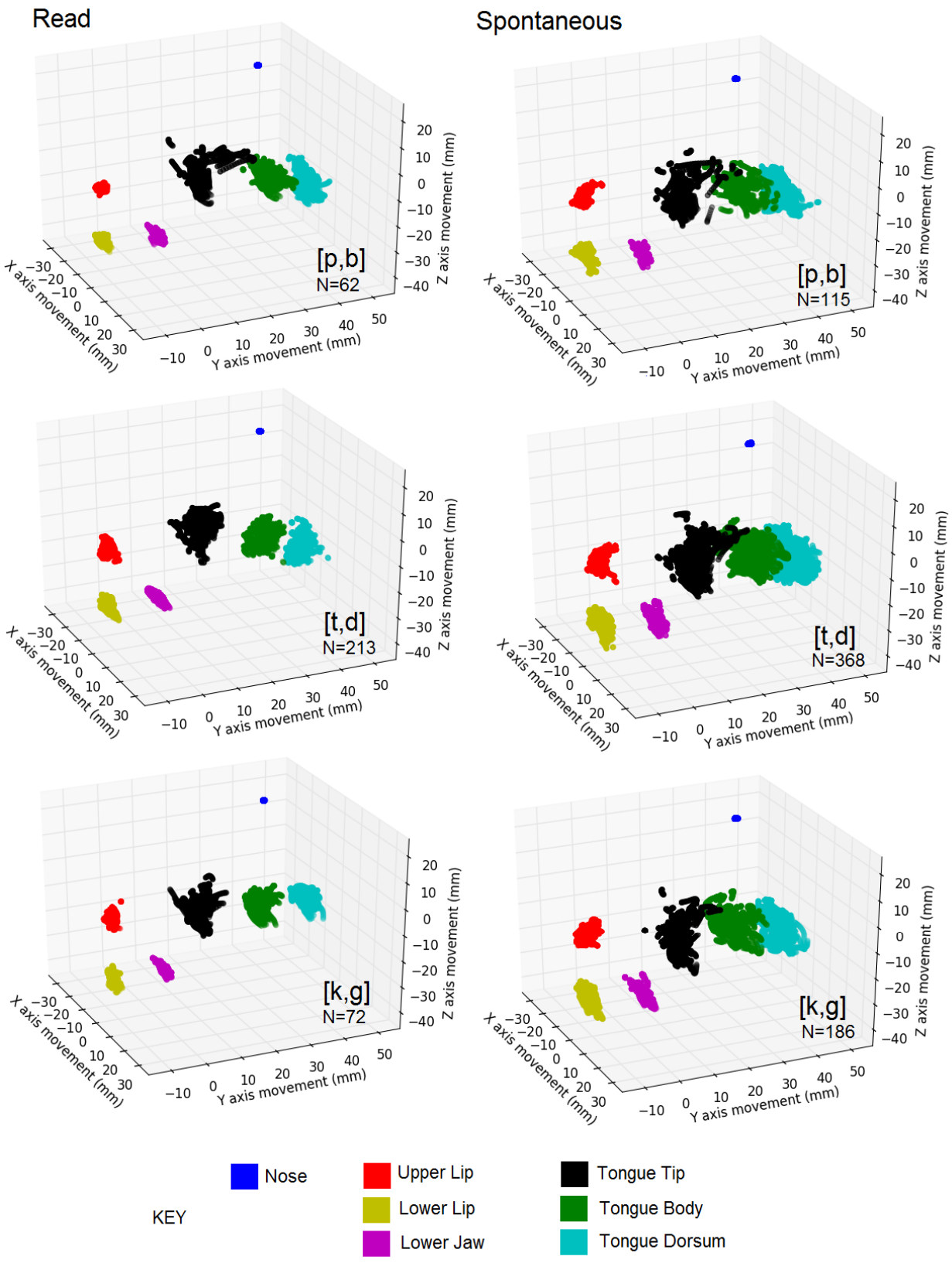
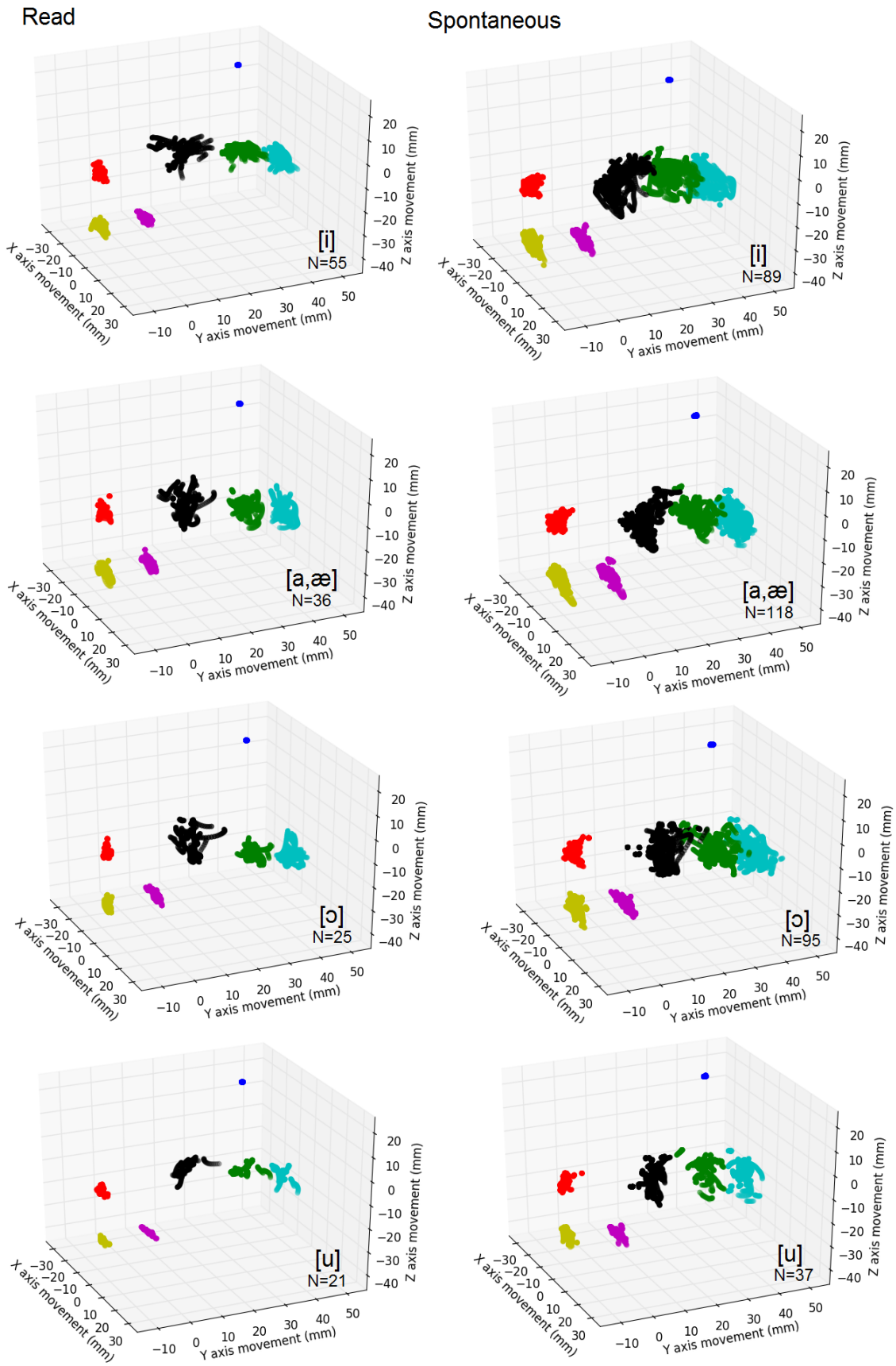


Figure 4.2: Articulator trajectories for speaker r33cs6 - Standard Southern British English, male


Read
Spontaneous
[i]
N=55
[i]
N=89
[a,æ]
N=36
[a,æ]
N=118
[ɔ]
N=25
[ɔ]
N=95
[u]
N=21
[u]
N=37
X axis movement (mm)
Y axis movement (mm)
Z axis movement (mm)

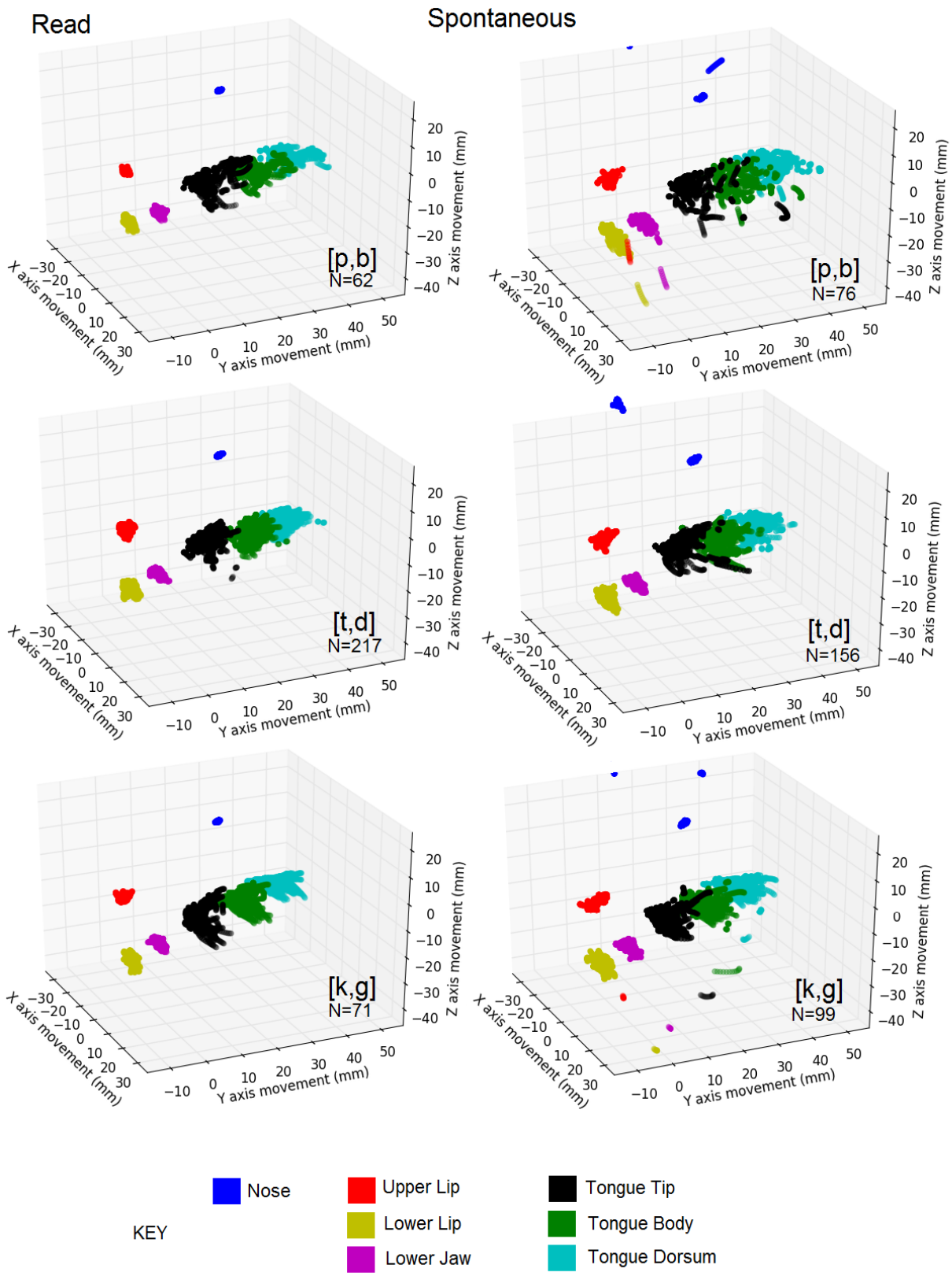
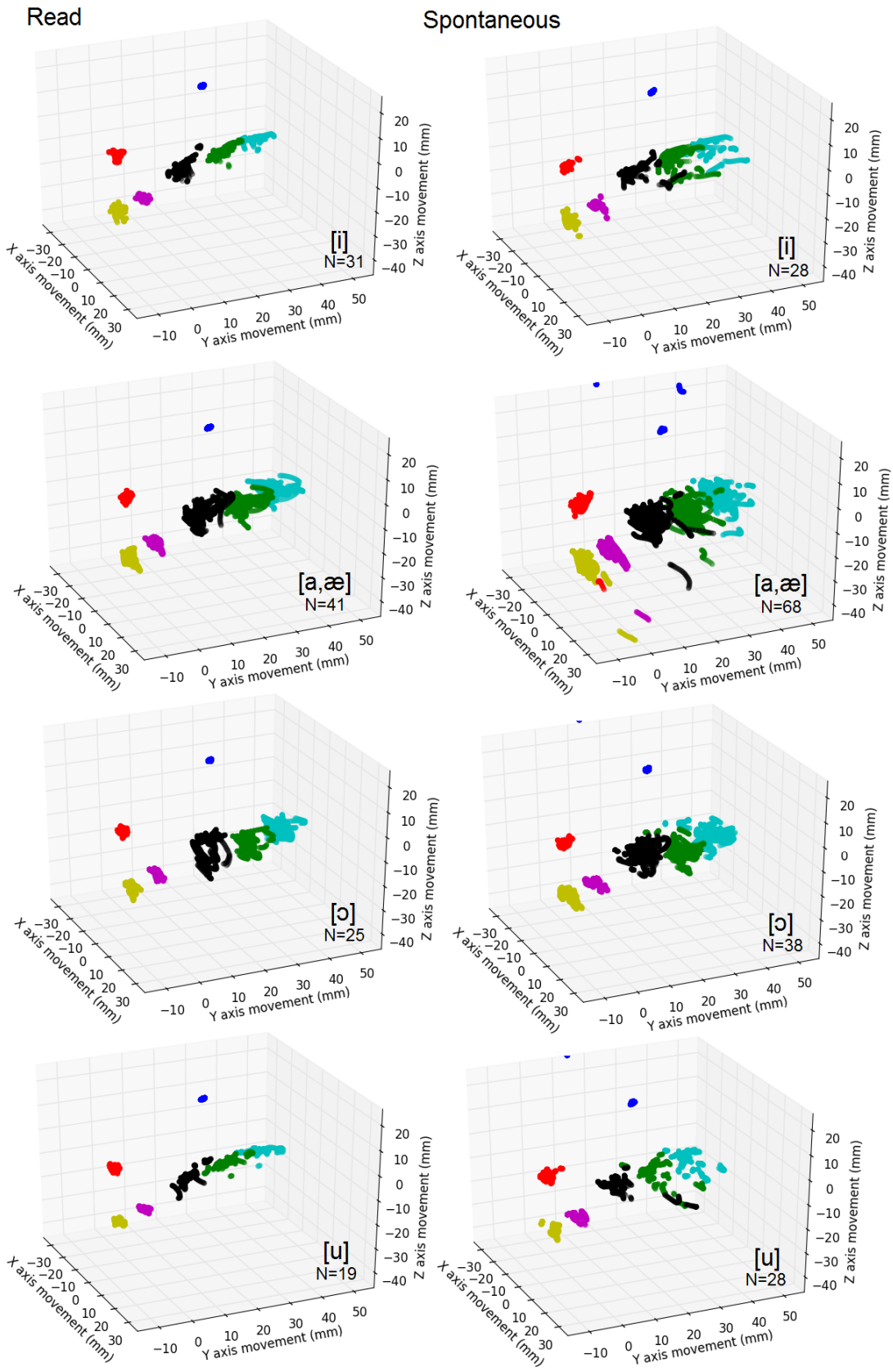


Figure 4.3: Articulator trajectories for speaker r34cs6 - Received Pronunciation, female


Read
Spontaneous
[i]
N=31
[i]
N=28
[a,æ]
N=41
[a,æ]
N=68
[ɔ]
N=25
[ɔ]
N=38
[u]
N=19
[u]
N=28
X axis movement (mm)
Y axis movement (mm)
Z axis movement (mm)

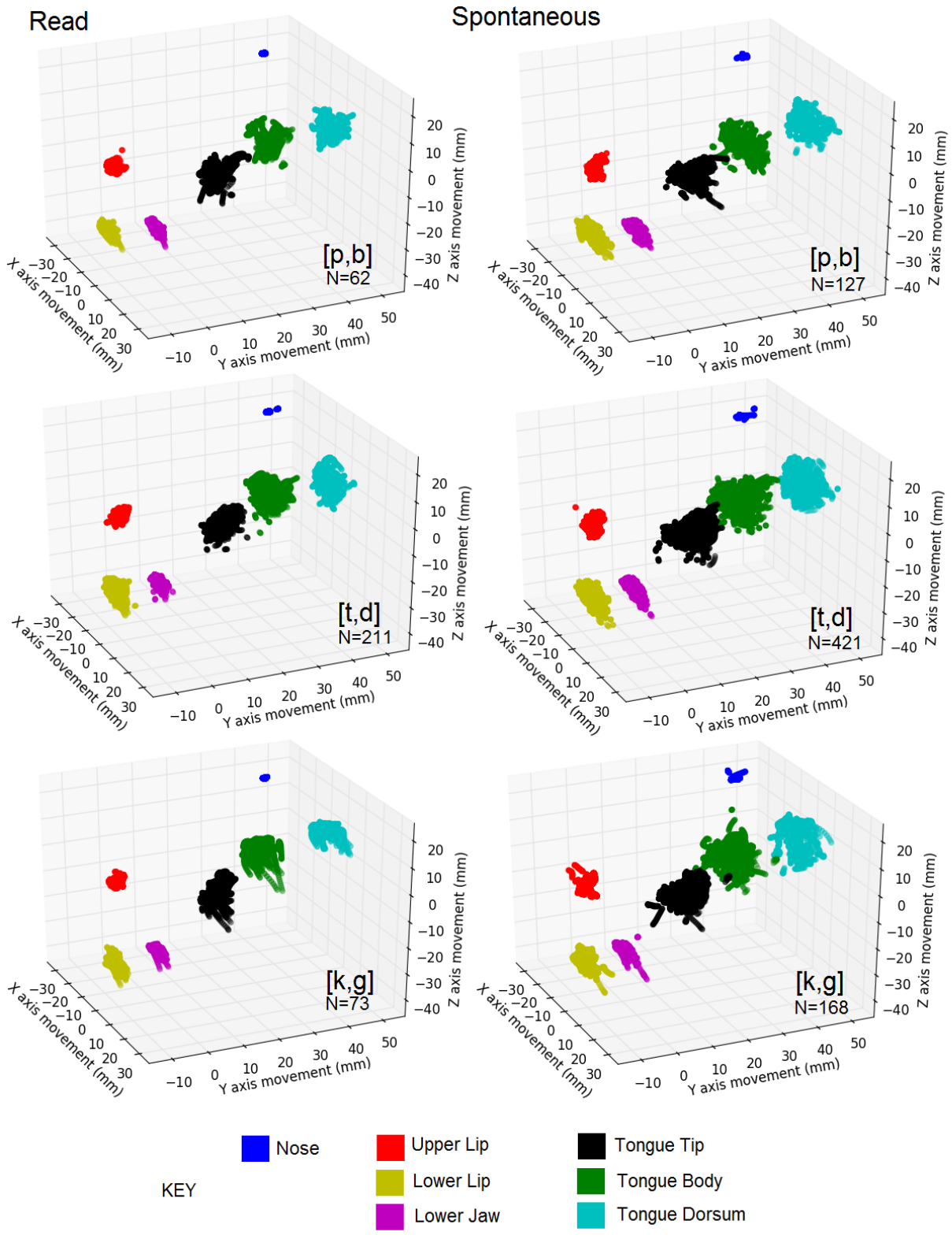
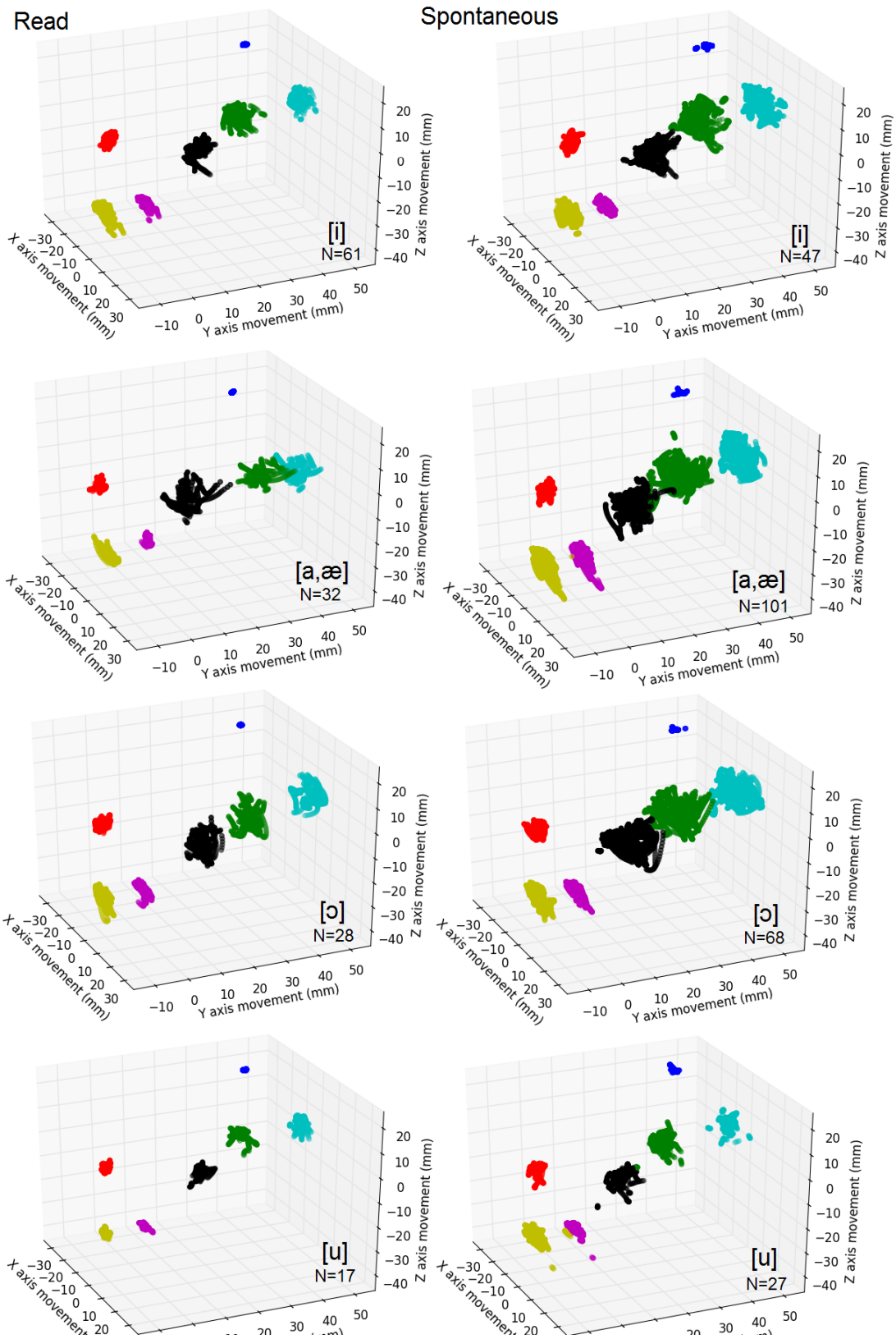


Figure 4.4: Articulator trajectories for speaker r35cs5 - Scottish Standard English, male


Read
Spontaneous
[i]
N=61
[i]
N=47
[a,æ]
N=32
[a,æ]
N=101
[ɔ]
N=28
[ɔ]
N=68
[u]
N=17
[u]
N=27
X axis movement (mm)
Y axis movement (mm)
Z axis movement (mm)

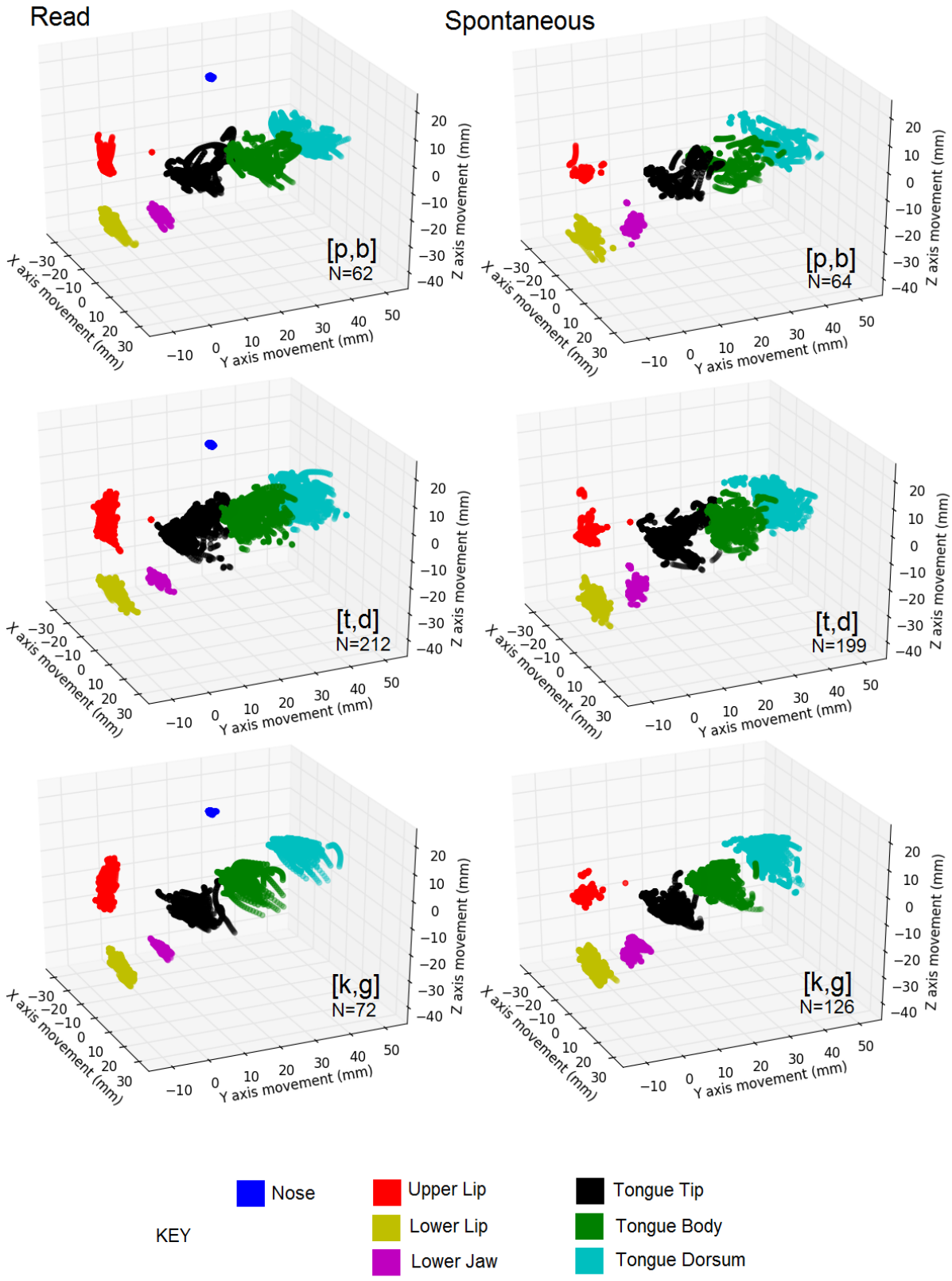
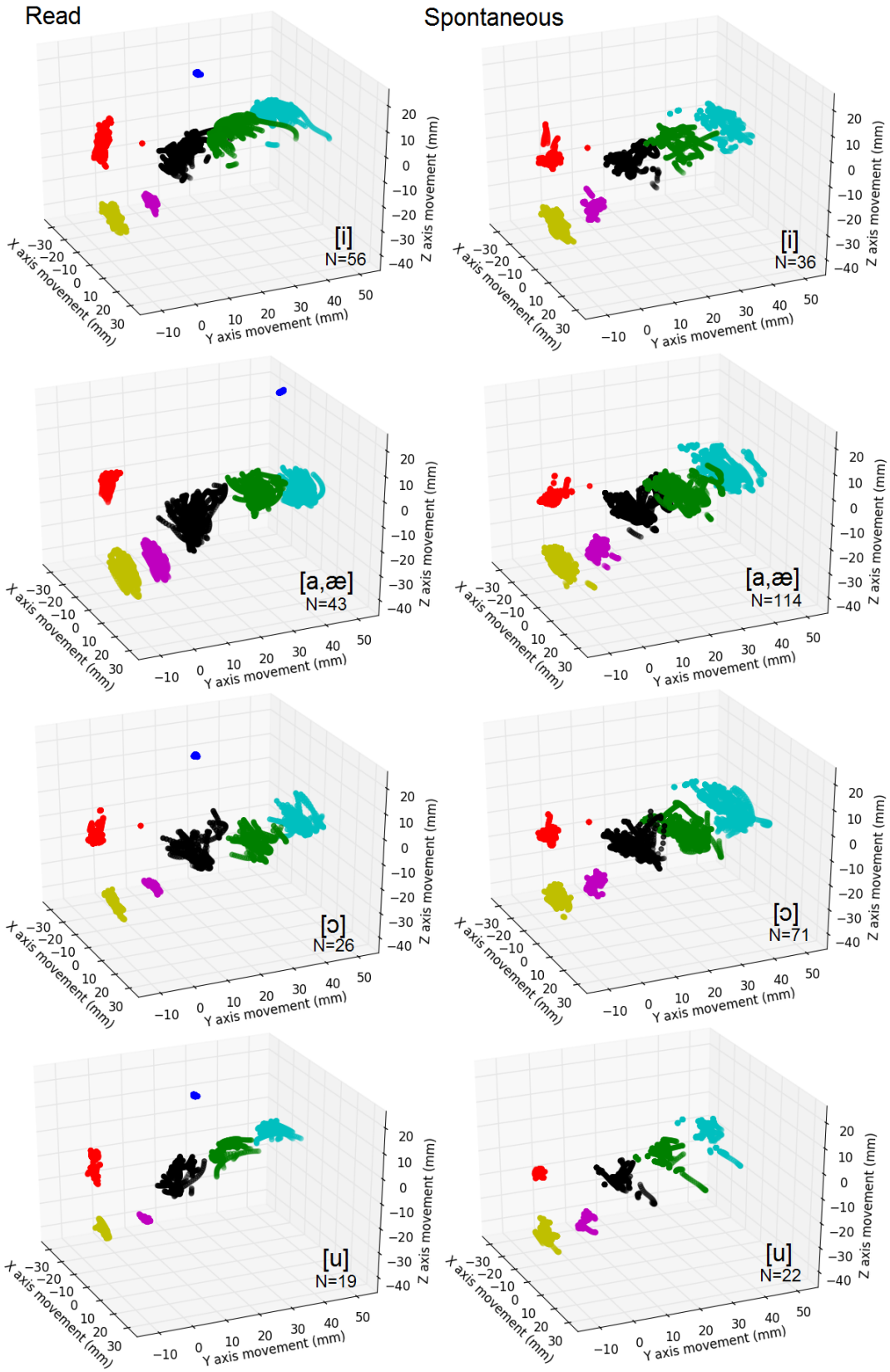


Figure 4.5: Articulator trajectories for speaker r35cs6 - Standard Southern British English, male


Read
Spontaneous
[i]
N=56
[i]
N=36
[a,æ]
N=43
[a,æ]
N=114
[ɔ]
N=26
[ɔ]
N=71
[u]
N=19
[u]
N=22
X axis movement (mm)
Y axis movement (mm)
Z axis movement (mm)

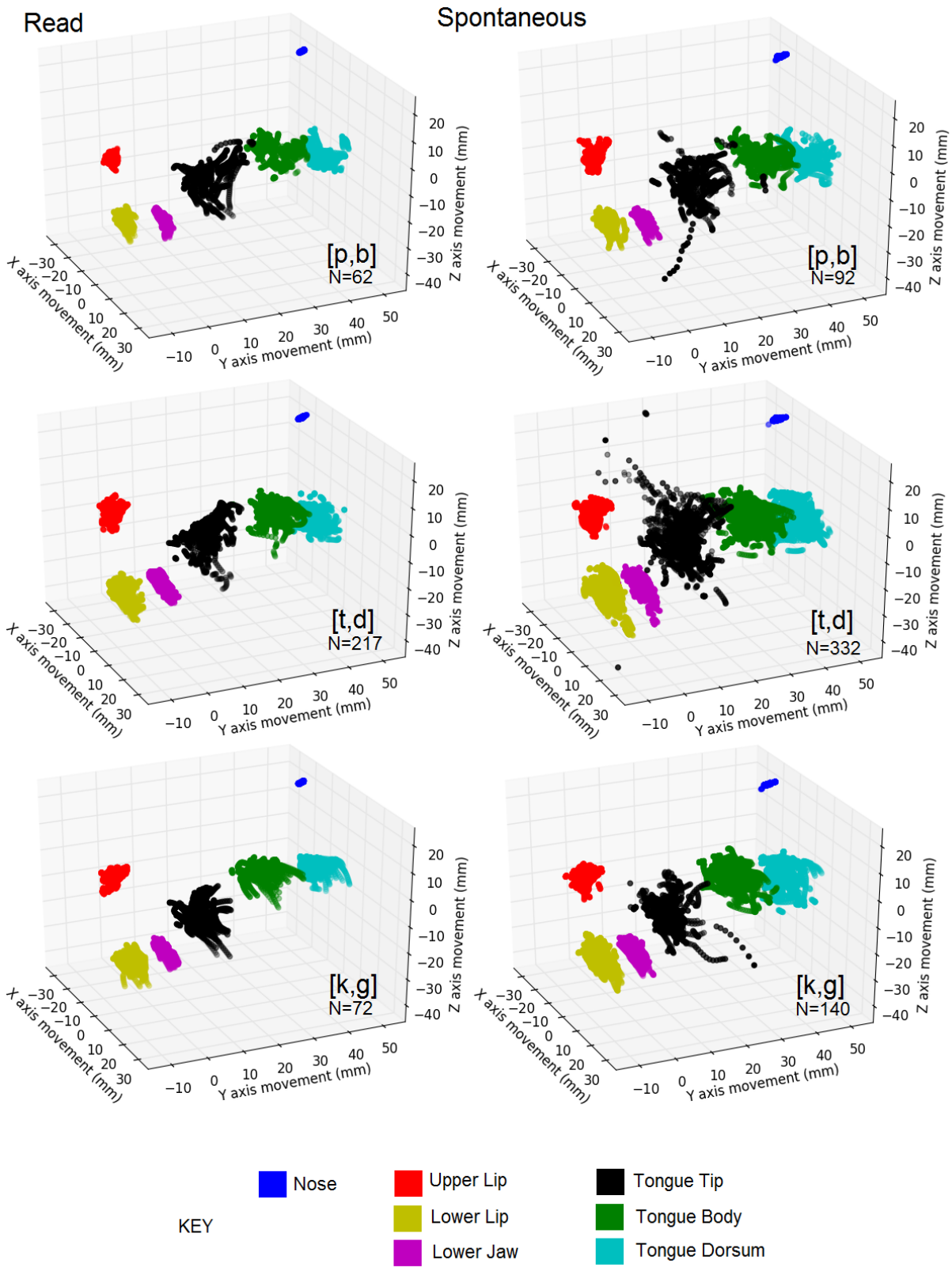
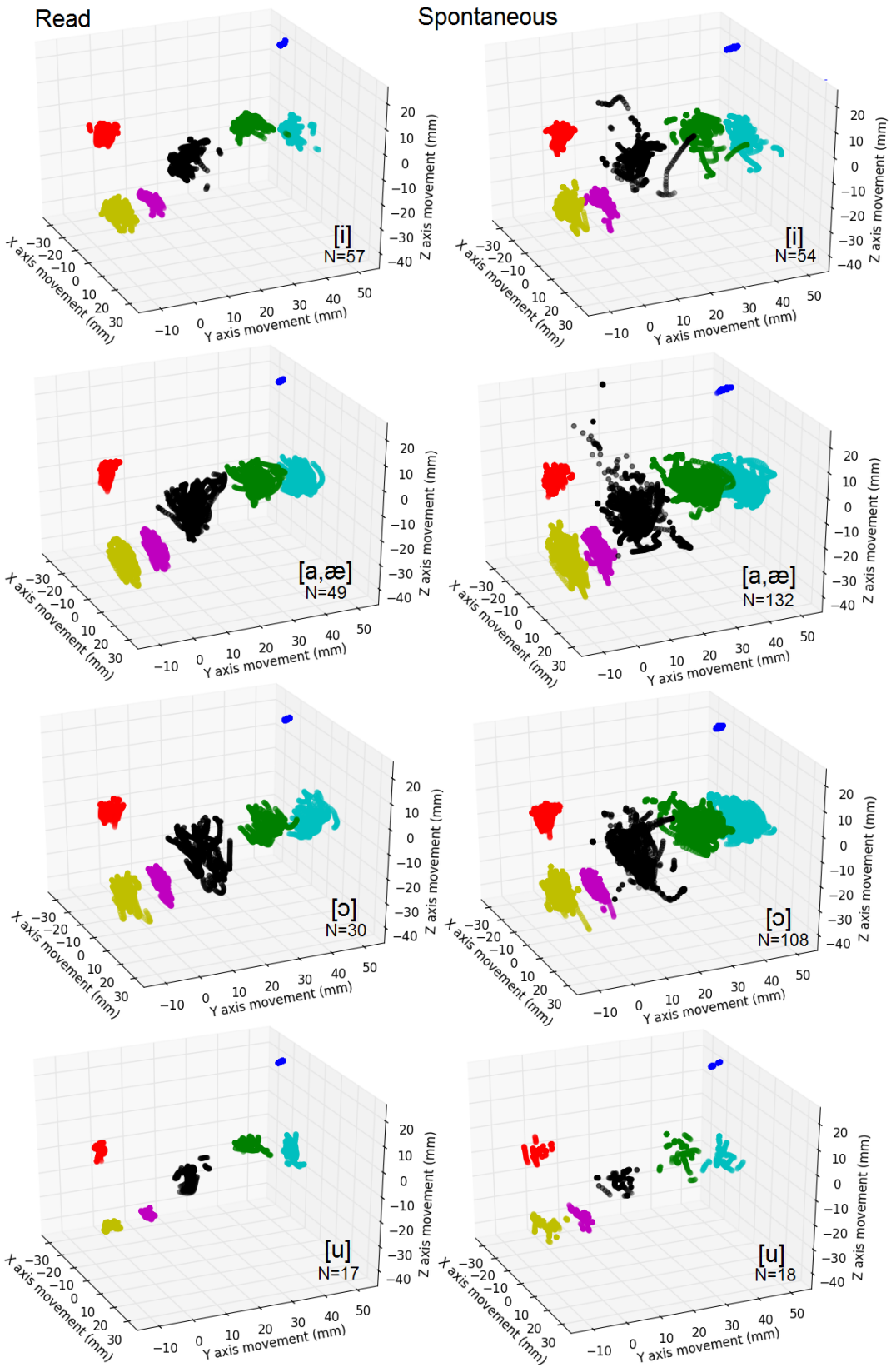


Figure 4.6: Articulator trajectories for speaker r36cs5 - Scottish Standard English, female


Read
Spontaneous
[i]
N=57
[i]
N=54
[a,æ]
N=49
[a,æ]
N=132
[ɔ]
N=30
[ɔ]
N=108
[u]
N=17
[u]
N=18
X axis movement (mm)
Y axis movement (mm)
Z axis movement (mm)

4.2 Variance

To analyse whether the articulators behaved significantly differently for a given phone, I performed paired t-tests on sensor position distributions at phone midpoints. Tables 4.1-4.7 display p-values for each phone's distribution of position for all occurring read and spontaneous tokens. Where there is a significant value, it may be inferred that there is a significantly different distribution of the given articulator's position between the two speech styles. I compare these findings with Jackson and Singampalli's (2009) critical articulator study. I also present boxplots of an articulator which significantly differs for each phone to be considered representative of the data. These plots are more abstract, but one-dimensional and more clear representations articulators in Figure 4.5.

Table 4.1: t-test p-values for the distribution of articulator position on the y and z axis for /p,b/

Sensor	Y axis	Z axis
LJ	0.101	0.958
UL	0.233	0.188
LL	0.017*	0.860
TT	0.001*	0.016*
TB	0.019*	0.359
TD	0.085	0.352

Table 4.2: t-test p-values for the distribution of articulator position on the y and z axis for /t,d/

Sensor	Y axis	Z axis
LJ	0.287	0.000*
UL	0.013*	0.000*
LL	0.002*	0.000*
TT	0.000*	0.000*
TB	0.051	0.000*
TD	0.305	0.005*

Table 4.3: t-test p-values for the distribution of articulator position on the y and z axis for /k,g/

Sensor	Y axis	Z axis
LJ	0.813	0.057
UL	0.607	0.000*
LL	0.120	0.054
TT	0.000*	0.013*
TB	0.301	0.111
TD	0.656	0.154

Table 4.4: t-test p-values for the distribution of articulator position on the y and z axis for /i/

Sensor	Y axis	Z axis
LJ	0.819	0.361
UL	0.673	0.003*
LL	0.189	0.093
TT	0.235	0.018*
TB	0.723	0.267
TD	0.812	0.886

Table 4.5: t-test p-values for the distribution of articulator position on the y and z axis for /æ,a/

Sensor	Y axis	Z axis
LJ	0.003*	0.405
UL	0.453	0.000*
LL	0.000*	0.207
TT	0.118	0.008*
TB	0.001*	0.251
TD	0.000*	0.321

Table 4.6: t-test p-values for the distribution of articulator position on the y and z axis for /ɔ/

Sensor	Y axis	Z axis
LJ	0.009*	0.898
UL	0.101	0.287
LL	0.002*	0.075
TT	0.278	0.861
TB	0.006*	0.635
TD	0.000*	0.877

Table 4.7: t-test p-values for the distribution of articulator position on the y and z axis for /u/

Sensor	Y axis	Z axis
LJ	0.119	0.101
UL	0.526	0.000*
LL	0.367	0.118
TT	0.766	0.787
TB	0.678	0.114
TD	0.418	0.207

For /p,b/ (Table 4.2, there were significantly different distributions for LL-y, TT-y, TT-z, and TB-y. This matches the identification of LL-y as a critical articulator, but not UL-y , whose distributions did not significantly differ between read and spontaneous speech. Figure 4.7 shows the difference in LL-y distributions across all /p,b/ tokens for each speech style. This figure shows that the lower lips are more forward, or protruded

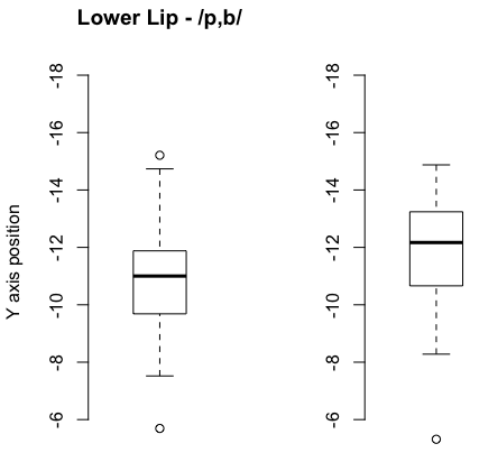


Figure 4.7: Midpoint distribution of LL-y for /p,b/ (Left - read speech, right - spontaneous speech)

in realisations of /p/ and /b/.

For /t,d/, there were significantly different distributions for LJ-z, UL-y, UL-z, LL-y, LL-z, TT-y, TT-z, TB-z, and TD-z. This means that all longitudonal distributions significantly differed between read and spontaneous tokens of /t/ and /d/. This compares with TT-y and LJ-y being considered critical for articulating /t,d/. Figure 4.8 shows the difference in TT-y distributions across all /t,d/ tokens for each speech style. The tongue tip is shown to be more retracted in this plot. There were many more tokens for /t,d/ than any other phone, so there may be a connection between amount of data and the number of articulator distributions which are significant.

For /k,g/, there were significantly different longitudonal distributions for UL-z, TT-y, and TT-z. This compares with TD-y as a critical articulator. Figure 4.9 shows the difference in UL-z distributions across all /k,g/ tokens for each speech style. The upper lip height varies much less, and is lower in spontaneous speech.

For /i/, there were significantly different distributions for UL-z and TT-z. TT-z is also considered a critical articulator for /i/. Figure 4.10 shows the difference in tongue tip height distributions across all /i/ tokens for each speech style. The tongue tip is generally lower in spontaneous speech. This infers that the tongue tip is not as close to reaching its articulatory target, close to the alveolar ridge (Browman & Goldstein, 1992).

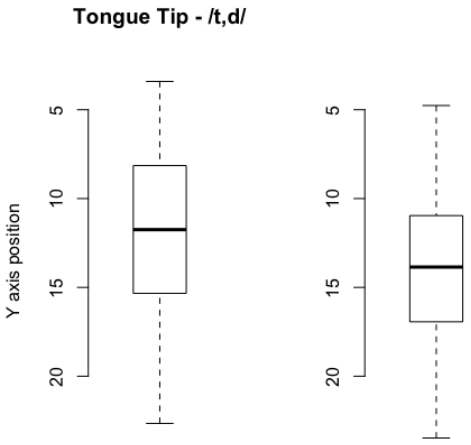


Figure 4.8: Midpoint distribution of TT-y for /t,d/ (Left - read speech, right - spontaneous speech)

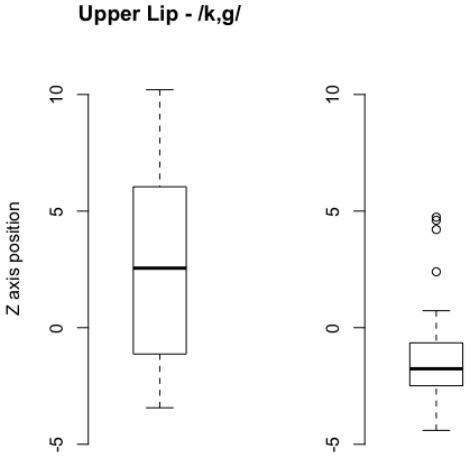


Figure 4.9: Midpoint distribution of UL-z for /k,g/ (Left - read speech, right - spontaneous speech)

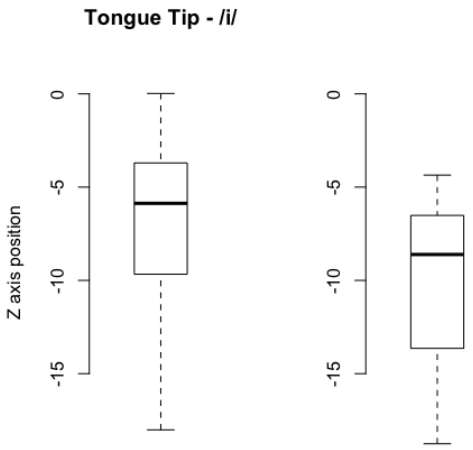


Figure 4.10: Midpoint distribution of TT-z for /t,d/ (Left - read speech, right - spontaneous speech)

For /æ/, there were significantly different distributions for LL-y, TT-y, TD-y, and TD-z. LL-y is also considered a critical articulator for /æ/. Figure 4.11 shows that the tongue back is significantly higher in spontaneous speech. As there are many outliers, the tongue dorsum's position may also be considered more variable in spontaneous speech.

For /ɔ/, there were significantly different distributions for LL-y, TB-y, and TD-y. TB-y is also considered a critical articulator for /ɔ/. Figure 4.12 shows that the tongue dorsum is generally further forward for spontaneous speech.

For /u/, there was a significantly different distribution of ULz. Jackson and Singampalli (2009) found no individual articulator to be critical. Figure 4.13 shows that the upper lip is less variable and much lower in spontaneous speech. The lack of variation in spontaneous data may be the result of having a low number of tokens (22) in the data.

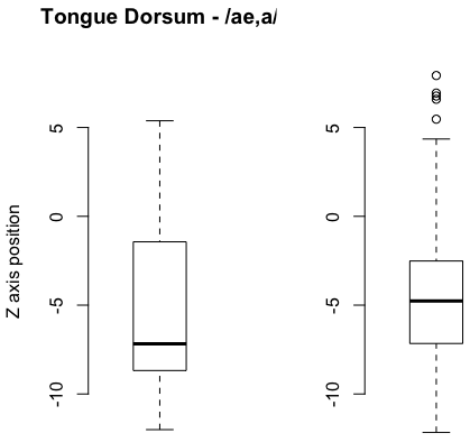


Figure 4.11: Midpoint distribution of TD-z for /æ/ (Left - read speech, right - spontaneous speech)

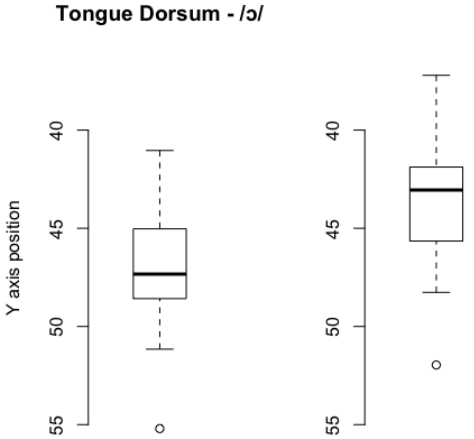


Figure 4.12: Midpoint distribution of TD-y for /ɔ/ (Left - read speech, right - spontaneous speech)

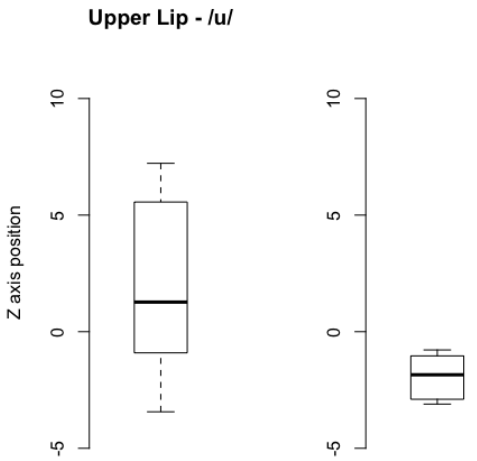


Figure 4.13: Midpoint distribution of UL-y for /u/ (Left - read speech, right - spontaneous speech)

4.3 Velocity

Parsons found that average articulatory speed was 20.452mm/s in read speech, and 17.544mm/s in spontaneous speech for all participants and across all of the production data (Parsons, 2015).

4.4 Summary

The data presented provides additional evidence that the articulation of read and spontaneous speech data is fundamentally different. Spontaneous speech is more articulatorily variable across all articulators and phones, and articulator speed is slower (Parsons, 2015). However, certain articulators differ more predictably for a given phone, shown by paired t-tests. A summary of articulators which significantly differed on two movement planes is shown in Table 4.8.

I also pointed out interspeaker differences on a qualitative level, including x-axis movement in spontaneous speech and global articulator configuration. These areas, as well as expanding a quantitative analysis to all six participants, require further investigation. In addition, critical articulators are not the only articulators to significantly differ by speech style. This means that variation does not only occur in articulators which are crucial to the perception of a phone.

Table 4.8: Summary of results - sensor coils which differed significantly for each phone

Phone	Significantly variable front-back (y-axis) position	Significantly variable longitudonal (z-axis position)
/p,b/	LL, TT, TB	TT
/t,d/	UL, LL	LJ, UL, LL, TT, TB, TD
/k,g/	TT	UL, TT
/i/		UL, TT
/æ/	LL, TT, TD	TD
/ɔ/	LL, TB, TD	
/u/		UL

Chapter 5

Discussion

This investigation has returned a vast quantity of information on the differences between read and spontaneous speech in terms of articulation. It has confirmed Parsons' (2015) observations, and answered the research question, of articulatory variation between speech style across seven vowels and consonants.

5.1 Implication in vocal motor theories

The AP/TD approach, where the production of phones is linked to constellations of articulatory gestures (Browman & Goldstein, 1992), does not sufficiently account for all observations made on this EMA data. For example, the gestures which are considered contrastive for /p/ and /b/ in AP/TD are limited to the degree of closure in the lips (Browman & Goldstein, 1992, p. 158). However, the distribution of front-back movement of the tongue, longitudonal movement of the tongue tip, and lip protrusion, significantly differ between read and spontaneous data. In addition, the tongue tip's distribution appears to be variable in both y and z-axis movement across most of the consonant and vowel phones studied. Vowels are specified in AP/TD by the constriction of the tongue tip and tongue body. However, evidence both from the current data and critical articulation (Jackson & Singampalli, 2009) show that the tongue configuration is not responsible for differences in /u/ production, nor is it critical for articulating /u/ or /æ/ in English EMA data.

Further, Browman and Goldstein (1992) suggest that when there are temporal and spatial constraints, phonetic gestures will shrink, but retain their specification of articulatory components. Temporal and spatial constraints align with spontaneous speech

in comparison with read speech which is non-planned, has increased levels of coarticulation (Nakamura et al., 2008) and is more error-prone (Pouplier, 2007). The present data do not align with this aspect of AP/TD. For every phone analysed, there has been at least one articulator in each direction of spatital movement that differed in its positional distribution during production. As for the temporal factor, these articulatory differences occur between speech styles which substantially differ in articulator velocity (Parsons, 2015), where the articulators move on average 2.91mm/s faster in read speech. The observation that spontaneous speech utilises a larger area of the articulatory space, but has slower-moving articulators, seems counter-intuitive. There appears to be less gestural overlap in read speech, where articulator distributions are qualitatively more 'precise' (Parsons, 2015). The shortcomings that AP/TD has in explaining these data may be due to its focus on speech production, and not speech planning and the hearer's intended perception.

5.1.1 Incorporating speech planning

The DIVA model (Perkell et al., 2000; Tourville & Guenther, 2011) takes into account the physiological constraints of the 'biomechanical' vocal motor system, the psychological constraints of time in speech planning, and the sensory linkage (e.g. Munhall, Gribble, Sacco, & Ward, 1996) in human speech perception. The 'internal model' is the cognitive control unit for speech within the DIVA framework. The internal model contains information about the constraints of the speech articulators, including temporal properties, and the current state of the muscles used in speech production. This means that feedforward information is always available to a speaker during spontaneous speech (Perkell, 2012). In comparison, read speech should also have acoustic feedback and additional speech planning at its disposal. In other words, additional strategies of reaching an acoustic goal region for a given phone (Perkell et al., 2000) are possible in read speech, as well as being heavily dependent on the current configuration of the articulators. Qualitatively, the 3D scatter plots seem to follow this analysis. The regions for both speech styles in the production of /t,d/, where there is the highest number of tokens for any phone amongst all participants, are highly variable. Perkell suggests the reasons for this are that, contrary to AP/TD, the goal regions of phones are grounded in neurophysiology (Perkell, 2012) rather than specific locations and specifications of individual articulators.

In terms of differences between read and spontaneous speech, theories incorporating speech perception and the constraints of the vocal motor system more adequately account for the variation between a given phone, articulator, or DoubleTalk participant. There is a different type of information available in speech production during the read passages through acoustic feedback and speech planning.

5.1.2 Naturalness of laboratory speech

As an aside, there is debate concerning the fidelity of speech collected under laboratory conditions, no matter the considered spontaneity of the speech task (Scobbie, Stuart-Smith, Warner, Warren, & Hay, 2012). It is impossible for participants in EMA data collection to be left completely unmonitored by researchers. Observer effects (c.f., Labov, 1972) were mitigated as far as possible by collecting DoubleTalk data in dyads after familiarisation with the lab environment (Scobbie et al., 2013). The comparisons made between reading the *Comma Gets a Cure passage*, and spontaneous responses to a previously-unseen spot the difference task were considered the most disparate tasks in terms of speech style during the composition of this investigation. For a corpus to reflect natural speech for use in speech technology, this seems sufficient in the current state of these systems. However, for psycholinguistic and sociolinguistic studies, the perceived naturalness of these laboratory tasks might not be sufficient (Scobbie et al., 2012; King et al., 2007).

5.2 Relationship with Critical Articulation

The relationship between speech style articulator configuration and critical articulation does not seem straightforward. The general trend in the present data is that articulators considered critical to articulate a given phone are not the articulators which vary in configuration between speech styles. Out of the ten articulators considered critical for the phones analysed in this investigation (Jackson & Singampalli, 2009), only four also significantly differed between speech styles, and 22 additional articulators' movements were identified as differentiating read and spontaneous speech. Because certain articulators are crucial for the perception of a phone category, they may not differ in specification between speech styles so that the perception of the phone remains the same.

Critical articulators rely on statistical correlations between an articulator and a

phone category (Kim et al., 2014), for example specified in the IPA framework. Between speech styles, and in speech variation more generally, different articulators are more sensitive to this varaition. This was shown by significant differences in articulator movements, most commonly in the tongue tip in all of the consonant phones, /i/ and /æ/, meaning that these articulators predictably differ between style.

One aspect of physiologically-real speech that critical articulator studies miss is aspects of coarticulation, crucial to insights into speech style and rate variation. The phonological categories to which critical articulators are assigned are abstract entities, and their ordering is assumed to be linear (Kim et al., 2014). This approach is more reliable for mathematically modelling speech (Felps et al., 2010), and more consistent with AP/TD (Jackson & Singampalli, 2009), but does not faithfully reflect speech style differences or the human vocal motor system.

5.3 Practical applications

Incorporating variable data, such as the current findings, into articulatory speech synthesis and recognition is challenging but necessary to answer current research problems. These problems include implementing more natural transitions between prosodic boundaries (Farrús, Lai, & Moore, 2016), and recognising natural speech with its disfluent and error-prone nature. By closely investigating differences between individual phones, the features and their weighting used in acoustic modelling for speech synthesis and ASR can become more faithful to both speech style and speech production in general, and hopefully reduce error rates in these systems. There has so far been mixed success in using articulatory data for acoustic features and landmarks, and ambitious systems using real-time speech kinematics (King et al., 2007)

In terms of acoustic-to-articulatory inversion, having greater insight into how articulatory patterning varies allows better learning of the relationship between articulation and the acoustic signal. A better understanding of the speech articulators allows better parameter tuning of HMMs or NNs for inversion (Papcun et al., 1992). Felps and colleagues (2010) have already compiled lists of articulators which are to be given higher feature weights in Mel Frequency Cepstral Coefficients used in TTS and ASR. Here, critical articulators were used for inversion. Further investigation into speech style would help fine tune these weights.

5.4 Further work and concluding remarks

There is scope to greatly expand the findings of the present investigation. Key areas to gain insight would include collecting more phones for analysis, quantitatively analysing individual phones and articulators for the other five participants, and finding other ways of presenting this complex data. Pinpointing individual articulator trajectory start and endpoints, as well as taking measurements of velocity for individual sensors and phones. Wieling and colleagues measured normalised and averaged articulator trajectories in their investigation on Dutch dialectal differences (2016). They plotted average trajectories of individual articulators producing individual phones which, in turn, overlaid the entire distribution of articulator movement. This more abstract representation may more clearly reflect differences between speech style, and can be normalised across many participants. Alternative measurements for articulator trajectories are also possible. Articulator speed for individual phones may be calculated for *.pos* files from the equation in Parsons (2015, p. 21). Another method, utilising Articulate Assistant Advanced (Wrench, 2007) software, involves calculating 'tangential velocity'. This metric measures the height of an articulator, the tongue tip in previous investigations (Purse, Turk, & Fruehwald, 2016), relative to local minima - assumed to be a resting point of an articulator - along the time axis. This metric could be incorporated to degree of protrusion of the lip articulators, tongue frontness, among others.

It is feasible to couple critical articualtor studies with investigations into speech style and planning. This would answer questions raised by this investigation about the relationship between articulatory variation and critical articualtion. Similar recent studies have linked critical articulators to 'emotional goals' (Kim et al., 2015) for use in speech technology. It is hoped that a critical articualtor investigation into speech style could identify similar markers in speech kinematics to identify and synthesise natural speech more faithfully.

Additional interesting areas of investigation may include analysing speech errors within the DoubleTalk data. Studying speech errors and acquisition has uncovered countless phonetic and articulatory phenomena invisible to speech perception and acoustic data (Pouplier, 2007). Therefore, studying errors will allow more rigorous analysis of the DIVA model (Perkell, 2012) which seemed to best explain read and spontaneous differences in the present data. Due to the nature of phrase-level tran-

scriptions, this type of investigation would rely on a time-consuming process of using the FAVE-align toolkit. However, it would be worthwhile to expand this investigation across all DoubleTalk files - for as many of the participants and ARPAbet phones as is possible without errorful data. Combining and normalising measurements across participants will allow a more concrete analysis of variation in articulatory movements, and the patterns observed in this study may become more clear.

5.4.1 Conclusion

The scope of this investigation has been mostly exploratory, but has confirmed clear quantitative and qualitative differences in the articulation of read passages and a spontaneous speech task. There is a tangible need for further investigation incorporating more speakers, speech tasks, phones, and combinatory approaches - all of which are possible with DoubleTalk data.

Appendix A

posplot.py

```
import numpy as np
import sys
# !r35cs5sponUW.png is currently missing
fname='r35cs5comma.pos'
data = np.fromfile(fname,dtype=np.float32).reshape(-1, 12, 7)
# fname2='r35cs5std2.pos'
# data2 = np.fromfile(fname2,dtype=np.float32).reshape(-1, 12, 7)

k1 = data[1720:1726,:,:]
k2 = data[2682:2688,:,:]
k3 = data[2836:2850,:,:]
k4 = data[3194:3204,:,:]
k5 = data[3946:3958,:,:]

k_data = np.concatenate((k1,k2,k3,k4,k5),axis=0)

print data.shape
print k_data.shape

# SELECTION of [frame, EMA articulator, coordinate]

# Lower Jaw
ljxdata = k_data[:,6,0]
```

```
ljydata = k_data[:,6,1]
ljzdata = k_data[:,6,2]
# Upper Lip
ulxdata = k_data[:,3,0]
ulydata = k_data[:,3,1]
ulzdata = k_data[:,3,2]
# Lower Lip
llxdata = k_data[:,4,0]
llydata = k_data[:,4,1]
llzdata = k_data[:,4,2]
# Tongue tip
ttxdata = k_data[:,9,0]
ttydata = k_data[:,9,1]
ttzdata = k_data[:,9,2]
# Tongue body
tbxdata = k_data[:,8,0]
tbydata = k_data[:,8,1]
tbzdata = k_data[:,8,2]
# Tongue dorsum
tdxdata = k_data[:,7,0]
tdydata = k_data[:,7,1]
tdzdata = k_data[:,7,2]
# Nose
nexdata = k_data[:,2,0]
neydata = k_data[:,2,1]
nezdata = k_data[:,2,2]

from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt
import matplotlib.markers

fig = plt.figure() # Initialise
fig.suptitle('')
```

```
ax = fig.add_subplot(111,projection='3d')

# layers and formatting of plot
ax.scatter(ttxdata,ttydata,ttzdata,color='k',marker='o') # T tip
ax.scatter(tbxdata,tbydata,tbzdata,color='g',marker='o') # T body
ax.scatter(tdxdata,tdydata,tdzdata,color='c',marker='o') # T dorsum
ax.scatter(ulxdata,ulydata,ulzdata,color='r',marker='o') # U lip
ax.scatter(llxdata,llydata,llzdata,color='y',marker='o') # L lip
ax.scatter(ljxdata,ljydata,ljzdata,color='m',marker='o') # L jaw
ax.scatter(nezdata,neydata,nezdata,color='b',marker='o') # Nose (ref)

ax.set_xlabel('X axis movement (mm)') # Axis labels (lims)
ax.set_ylabel('Y axis movement (mm)')
ax.set_zlabel('Z axis movement (mm)')


ax.view_init(elev=28,azim=-24) # Plot viewpoint (if not default)
ax.set_zlim(-45,25)
plt.xlim(-35, 35)
plt.ylim(-15,55)
# plt.zlim(-45,25)
plt.show()
```

Appendix B

framewav.py

```
import praatio
from os.path import join
from praatio import tgio
import numpy as np

# Open desired Textgrid, find tier names
tg = tgio.openTextgrid("r35cs5comma.TextGrid")
print tg.tierNameList
#
## Get all instances of a phone (N)
## Tierdict is always "speech - phone" from FAVE
# k_list = tg.tierDict["speech - phone"].find("N")
# print k_list
# print len(k_list)

### PLOTTING SECTION ###

# All timestamps for one phone
tegd = tgio.openTextgrid("r35cs5comma.TextGrid")
tier = tegd.tierDict["silences - phone"]

posname ='r35cs5comma.pos'
posdata = np.fromfile(posname,dtype=np.float32).reshape(-1, 12, 7)
```

```
lenframe = posdata.shape[0] # entire file length in frames
lenwav = tegd.maxTimestamp # entire file length in seconds
counter = 0
for start, stop, label in tier.entryList:
    # counter += 1 # put inside if loop for order of K's only
    start = start / lenwav
    stop = stop / lenwav
    newstart = lenframe * start
    newstop = lenframe * stop
    if label == "TH" or label == "DH":
        counter += 1
        print("k%d = data[%d:%d,:,:]" % (counter, newstart, newstop))
        # prints the correct format for plotting
```

References

Boersma, P., & Weenink, D. (2005). Praat (version 4.3. 31). *Amsterdam: University of Amsterdam. Online: http://www. fon. hum. uva. nl/praat*.

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

Cho, T. (2006). Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in english. *Laboratory phonology*, *8*, 519–548.

Coupland, N. (1980). Style-shifting in a cardiff work-setting. *Language in Society*, *9*(1), 1–12.

Ellis, L., & Hardcastle, W. J. (2002). Categorical and gradient properties of assimilation in alveolar to velar sequences: evidence from EPG and EMA data. *Journal of Phonetics*, *30*(3), 373–396.

Farnetani, E., & Recasens, D. (2010). Coarticulation and connected speech processes. *The Handbook of Phonetic Sciences, Second Edition*, 316–352.

Farrús, M., Lai, C., & Moore, J. D. (2016). Paragraph-based prosodic cues for speech synthesis applications. *Barnes J, Brugos A, Shattuck-Hufnagel S, Veilleux N, editors. Speech Prosody 2016; 2016 May 31-June 3; Boston (MA, USA).[place unknown]: International Speech Communication Association; 2016. p. 1143-7.*.

Felps, D., Geng, C., Berger, M., Richmond, K., & Gutierrez-Osuna, R. (2010). Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database.

Geng, C., Turk, A., Scobbie, J. M., Macmartin, C., Hoole, P., Richmond, K., ... others (2013). Recording speech articulation in dialogue: Evaluating a synchronized double electromagnetic articulography setup. *Journal of Phonetics*, *41*(6), 421–431.

Gick, B., Wilson, I., & Derrick, D. (2012). *Articulatory phonetics*. John Wiley & Sons.

Grosjean, F., & Collins, M. (1979). Breathing, pausing and reading. *Phonetica*, *36*(2),

98–114.

Harrington, J., Fletcher, J., & Roberts, C. (1995). Coarticulation and the accented/unaccented distinction: evidence from jaw movement data. *Journal of Phonetics*, *23*(3), 305–322.

Iskarous, K., Shadle, C. H., & Proctor, M. I. (2011). Articulatory–acoustic kinematics: The production of american english/s. *The Journal of the Acoustical Society of America*, *129*(2), 944–954.

Jackson, P. J., & Singampalli, V. D. (2009). Statistical identification of articulation constraints in the production of speech. *Speech Communication*, *51*(8), 695–710.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). Pearson London.

Kim, J., Lee, S., & Narayanan, S. S. (2014). Estimation of the movement trajectories of non-crucial articulators based on the detection of crucial moments and physiological constraints. In *Fifteenth annual conference of the international speech communication association.*

Kim, J., Toutios, A., Lee, S., & Narayanan, S. S. (2015). A kinematic study of critical and non-critical articulators in emotional speech production. *The Journal of the Acoustical Society of America*, *137*(3), 1411–1429.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007). Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, *121*(2), 723–742.

Labov, W. (1972). *Sociolinguistic patterns* (No. 4). University of Pennsylvania Press.

Lawson, E., Stuart-Smith, J., Scobbie, J. M., Nakai, S., Beavan, D., Edmonds, F., ... others (2015). Seeing speech: an articulatory web resource for the study of phonetics [website].

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lindblom, B. (1986). Phonetic universals in vowel systems. *Experimental phonology*, 13–44.

Matthies, M., Perrier, P., Perkell, J. S., & Zandipour, M. (2001). Variation in anticipatory coarticulation with changes in clarity and rate. *Journal of Speech, Language, and Hearing Research*, *44*(2), 340–353.

McCullough, J., Somerville, B., & Honorof, D. (2000). Comma gets a cure. *International Dialects of English Archive*.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the mcgurk effect. *Perception & psychophysics*, *58*(3), 351–362.

Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, *22*(2), 171–184.

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., & Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *The Journal of the Acoustical Society of America*, *92*(2), 688–700.

Parsons, P. (2015). *Acoustic-to-Articulatory Inversion Using the DoubleTalk Corpus.* (unpublished thesis)

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics*, *25*(5), 382–407.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., & Jackson, M. T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, *92*(6), 3078–3096.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., ... Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, *28*(3), 233–272.

Pouplier, M. (2007). Tongue kinematics during utterances elicited with the slip technique. *Language and Speech*, *50*(3), 311–341.

Purse, R., Turk, A., & Fruehwald, J. (2016). '/t,d/ deletion': Articulatory gradience in variable phonology. *Proceedings of LabPhon15, Ithaca NY*.

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). Fave (forced alignment and vowel extraction) program suite. *URL http://fave. ling. upenn. edu*.

Scobbie, J. M., Stuart-Smith, J., Warner, N., Warren, P., & Hay, J. (2012). Experimental design and data collection. *The Oxford Handbook of Laboratory Phonology*.

Scobbie, J. M., Turk, A., Geng, C., King, S., Lickley, R., & Richmond, K. (2013). The edinburgh speech production facility doubletalk corpus. *Proceedings of 14th Interspeech, Lyon*.

Singampalli, V. D., & Jackson, P. J. (2007). Statistical identification of critical, dependent and redundant articulators. In *Interspeech 2007: 8th annual conference of*

the international speech communication association, vols 1-4 (pp. 2736–2739).

Steiner, I., Richmond, K., Marshall, I., & Gray, C. D. (2012). The magnetic resonance imaging subset of the mngu0 articulatory corpus. *The Journal of the Acoustical Society of America*, *131*(2), EL106–EL111.

Tourville, J. A., & Guenther, F. H. (2011). The diva model: A neural theory of speech acquisition and production. *Language and cognitive processes*, *26*(7), 952–981.

Tuller, B., Harris, K. S., & Kelso, J. S. (1982). Stress and rate: Differential transformations of articulation. *The journal of the Acoustical society of America*, *71*(6), 1534–1543.

Wells, J. C. (1982). *Accents of english, volume: 1, 2.* Cambridge University Press.

Westbury, J. R., Turner, G., & Dembowski, J. (1994). X-ray microbeam speech production database user's handbook [version 1.0].

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., ... Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, *59*, 122–143.

Wrench, A. A. (2000). A multi-channel/multi-speaker articulatory database for continuous speech recognition research. *Phonus.*, *5*, 1–13.

Wrench, A. A. (2007). Articulate assistant advanced user guide: Version 2.07. *Edinburgh: Articulate Instruments Ltd*.