Articulation and Coarticulation in the Lower Vocal Tract

Greg Kochanski and John Coleman, *University of Oxford*

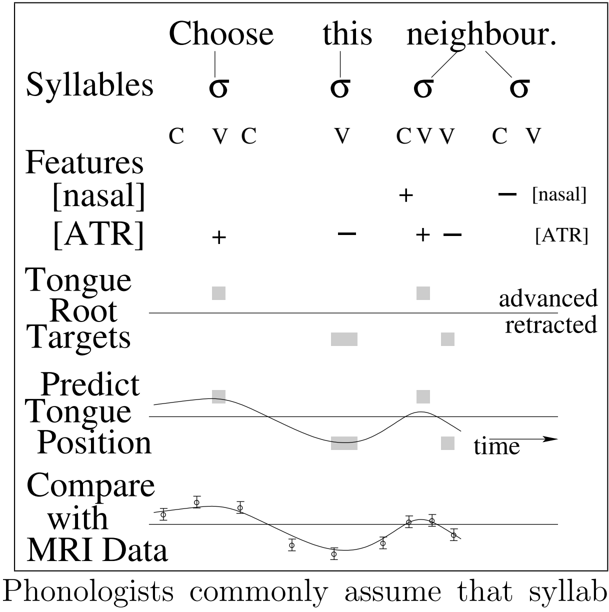
This document was a proposal to the UK's Economics and Social Research Council in 2004. It was accepted and was the basis for project RES-000-23-1094 which ran from June 2005 to March 2008.

1 Introduction

This proposal is aimed at understanding the strategies that are used to control articulators during speech and at clarifying the relationships between phonological features and articulatory motions. This experimental project involves the collection and analysis of Magnetic Resonance Imaging (MRI) and acoustic speech data, combined with numerical modelling of articulatory motions.

Movements of the tongue root and velum are very important in speech production. The velum position effectively defines nasal consonants; the advancement or retraction of the tongue root determines the volume of the pharynx, the largest cavity of the vocal tract. However, the back half of the vocal cavity has had relatively little attention. Access is difficult, and subjects do not tolerate many experimental techniques that work well for the front of the vocal cavity.

Schematic of this project.



Phonologists commonly assume that syllables, words and utterances are constructed from features, such as [+ATR] or [+nasal]. In our articulatory model, we assume that each feature maps onto an articulatory target. In Figure 1, we show simple targets specifying advanced/retracted tongue root as grey rectangles.

From such targets, we will compute articulatory trajectories, using established models of muscle control, *i.e.* Minimum Variance [Harris and Wolpert, 1998] and Minimum Jerk [Flash and Hogan, 1985], which have been extensively tested on eye and limb motions. These models can be adapted to speech using techniques in Kochanski et al. [2003].

We will then compare each computed trajectory to MRI measurements of the tongue root and velum. The targets and parameters in the muscle control models will be adjusted to obtain the best fit to our MRI corpus. Differences between the models and data will reflect both the consistency of the phonology with the speech and the correct choice of motor control strategy. This might lead to revisions of the phonological representations.

1.1 Goals

- Measure how well a phonological representation combined with a realistic motor control model can reproduce articulatory trajectories. We plan to test the hypothesis that phonological features control speech, which is critical for connecting phonology to phonetics and ultimately to biology. While it will clearly not fail entirely, there has been little in the way of quantitative evaluation, especially for the lower vocal tract.

- Determine whether Advanced Tongue Root, [±ATR], is a phonological feature in English, as proposed by Halle and Stevens [1969] (although challenged by Harshman et al. [1977]). Does its inclusion lead to better articulatory models than other phonological descriptions of equivalent complexity?

The following goals stand on their own, but they are also necessary intermediate steps to the above two goals:

- Experimentally explore the relationship between phonological features and articulatory targets. Is there a consistent association between features and motions?

- Find the optimal techniques for dynamic MR imaging of speech, including audio processing techniques to separate speech from MRI noise.

1.2 Articulatory Targets

Speech sounds are articulated in different ways, depending on context. For example, the /s/ in “soon” is produced with the lips rounded because of the requirement of the following vowel, whereas /s/ in “seen” has the lips retracted. Coarticulation can also reach beyond neighbouring sounds.

In the models we use, an “articulatory target” is the articulatory invariant of a phonological feature. It is a theoretical construct and may not be precisely realised, especially if there are other nearby and/or conflicting targets.

A target can be a combination of a desired articulator position and velocity, as when one throws darts. For a dart to reach its target, the hand and arm must move through certain positions at the right speeds. Trade-offs may occur between position and speed or between the motions of one joint (analogous to articulator) and another. We will use a mathematical definition of target that will allow trade-offs between different articulations that achieve the same sound.

2 Background

2.1 Phonological Representation of Speech

In current phonological theory, words are constructed out of features, the atomic units of the field. There are only a few kinds of features, and a minimal pair of words like “bad” and “mad” are considered to differ by a single feature. Features can be mapped onto articulatory actions: for instance, [+nasal] specifies that the velum should be lowered, as for /m/ in “mad”.

Normally, not all features are specified for every phone (vowel or consonant). Unspecified features have been dealt with in two ways: through phonological rules that copy specified features into their unspecified neighbours [Henke, 1966] or through interpolation rules that operate on stretches where a certain feature is not specified [Pierrehumbert, 1980] to yield articulator positions.

There are unresolved debates about the assignment of features to phones, most prominently to “schwa” (/ə/), which is particularly susceptible to coarticulatory effects. Browman and Goldstein [1992] argued that the articulation of /ə/ is not tightly specified (it is nearly “targetless”) and that variation between different versions of /ə/ can be largely explained by coarticulation. Others, such as Gick [2002] disagree.

2.2 Measurements of Articulator Positions

Velum movements have been studied with the Electromagnetic Articulograph (EMA) (*e.g.* Wrench [1999]). Other techniques include Electromyography (EMG), fiberscope, and velotrace [Benguerel et al., 1977a,b, Benguerel and Cowan, 1974]. All four procedures are invasive and sometimes not well tolerated; data from them are consequently limited. Furthermore, any invasive procedure will disturb the subjects’ speech, raising questions about the naturalness of the data. Velum and tongue root movements have also been studied using X-ray cinefluorography, though safety standards prohibit further use. Good images of the tongue surface can also be obtained using ultrasonography. However, the air space between the tongue and the upper surface of the vocal tract cannot be filmed in this way, so it is difficult to relate tongue shapes to their acoustic consequences.

To study coarticulation and dynamics, we will use MR images of moving articulators [Demolin et al., 2002, Narayanan et al., 2004]. Some MRI measurements have been done on the velum, notably Demolin et al. [1998], though those were done on extended vowels. Medically-driven developments have produced stroboscopic MRI sequences for cardiac imaging (*e.g.* Carr et al. [2001]). These sequences build up a composite image from data taken during a number of beats. They have been applied to speech by Foldvik et al. [1993] and Mathiak et al. [2000].

2.3 Experimental Studies of Coarticulation

Whalen [1990] conducted an experiment that controlled the letter-by-letter presentation of each sentence. He showed that if speakers do not know what they are about to say, coarticulation is minimal. Coarticulation therefore involves cognitive planning.

Cohn [1993] provided an example of the articulatory effect of an underspecified feature. She contrasted English and French, showing very different coarticulation of [+nasal] consonants. In English, the neighbouring vowels (typically) have unspecified nasality, while in French, vowels typically have either [+nasal] or [−nasal].

Coarticulation has also been studied beyond nearest-neighbour phones. Representative studies are Öhman [1966, 1967], Benguerel and Cowan [1974], Slater and Coleman [1996], Magen [1997], West [2000a], Coleman [2003], and Hawkins and Nguyen [2004]. Coarticulatory influences have been observed over domains of two syllables (or even further: *cf* Heid and Hawkins [2000]) before the controlling phone. These effects are noticeable, as West [2000b] found, because listeners can deduce a blanked-out phone from the coarticulatory changes it causes. Such long-range effects are particularly interesting because most models of coarticulation can accommodate long-domain effects only if there are no intervening phones that strongly specify a target.

2.4 Models of Speech Articulation and Coarticulation

Models of coarticulation have a long history, going back at least to Öhman [1966]. Generally, such models assume (or produce, via a dynamical system) articulatory gestures that have long tails. The overlap of these tails explains coarticulation [Fowler, 1980]. Although there is much literature in the field, the focus is on abstract dynamical systems which may have predictive power but tend not to explain the underlying mechanisms.

Browman and Goldstein [1986, 1989], successfully explained the (apparent) deletion of some word-final sounds with their "Task Dynamic" model: a weak gesture remains, but is not large enough to have a noticeable acoustic effect. Unfortunately, this model was never published in complete detail or tested on substantial corpora. Fujisaki [1983] developed a related (but less abstract) model of articulatory control which has been extensively applied to intonation. Recent work along this line includes King and Wrench [1999], Blackburn and Young [2000], Kaburagi and Honda [2001], and Deng [2004]. In most recent work, abstract equations that govern the articulators are learnt from a database.

2.5 Reductionist Models of Muscle Control

Outside speech science, motor control has also been studied. Relative to §2.4, such models are generally more closely related to the underlying physiology. They are often testable in ways beyond their ability to reproduce training data, such as prediction of EMG activity [Christou et al., 2003] or motion in the presence of an obstacle [Sabes et al., 1998].

These models sprang from simulations [Feldman, 1986] of neural connections [Liddell and Sherrington, 1925]. Current models, like "Minimum Jerk" [Flash and Hogan, 1985] can deal with larger-scale, strategic choices among possible trajectories. In these models, the brain plans a trajectory that attempts to minimise error from a target while maximising the smoothness of motion (see Wolpert [1997], and references therein).

The Minimum Variance model [Harris and Wolpert, 1998] is a recent, elegant movement model that is closer to physiology and may be more accurate than Minimum Jerk. Articulators are not perfectly controllable; the discrete nature of neural signals creates unavoidable errors. Minimum Variance assumes a trajectory that minimises the root-mean-square (RMS) error at a target. It rejects rapid, jerky trajectories because they turn out to be less accurate.

3 Experiments

We plan three experiments: the first two are steps toward the third. The first one establishes the phonological representation needed for the final model-building; the second establishes optimal data acquisition techniques.

1. Relating Features to Targets.

2. Evaluation of Real-Time *vs.* Stroboscopic Magnetic Resonance Imaging.

3. Tongue root and velum movement: from MRI to articulatory model.

3.1 Acoustic Experiment 1: Relating Features to Targets

Representing words *via* phonological features is well established. Features primarily represent contrasts between words of the language. A language, however, might control articulatory positions in places that are not contrastive (*e.g.* redundant features: Jakobson et al. [1952]). Consequently, the normal feature assignments may not capture *all* articulatory targets.

For instance, many English speakers round their lips for the consonants /ʃ/, /ʒ/, /ʧ/, /ʤ/ and /ɹ/ [Brown, 1981], even in words like "cheese", where the vowel is un-rounded. This is not contrastive in English: there is no pair of words that are distinguished by rounding of /ʃ/. The rounding is presumably due to an articulatory target, but these consonants are not normally assigned a [+round] feature.

The idea behind Experiment 1 is to make a systematic survey for similar examples. If there is an articulatory target, the corresponding acoustic properties should display relatively little variability across environments and *vice versa* [Keating, 1988]. Thus, we hope to test and validate the phonological representation before we use it to build articulatory models in §3.3. Any extra articulatory targets (*i.e.* without an associated feature) we may find will be as important to the articulatory models in Experiment 3 (§3.3) as targets that are used contrastively.

3.1.1 Design

Stimuli will be a list of sentences taken from a corpus of newswire text (lightly edited as necessary). This will ensure a reasonably natural reading style, avoiding boredom or over-articulation.

We will find phone sequences in the form AXB where X is the phone under study, and AB is the context. We will look for a set of contexts, $\mathbb{C}$, that covers a wide variety of features and has about ten members. Ideally, every context in $\mathbb{C}$ should be applied to each phone under study. We will choose $\mathbb{C}$ to maximise the amount of context shared by the phones under study, $\mathbb{X}$.

$\mathbb{X}$ will be chosen with Experiment 3 (§3.3) in mind. It will contain [±ATR] vowels, and [±nasal] consonants (see Appendix B).

Given the variation in articulatory strategies from person to person (*e.g.* Lieberman [1976]), the minimum reasonable group size is about five speakers. We will collect both oral and nasal data to measure nasality.

3.1.2 Analysis

To learn and apply the signatures of articulatory targets, we will use classifier techniques developed in our current ESRC project [Coleman and Grabe, 2003]. As input to the classifier, we compute a feature vector[1] that will represent the variance in the power spectrum of phone X caused by changes in its environment. (Alternatively, the feature vector might include nonlocal measures and/or psychophysical quantities like those of O'Mard [2004] or Patterson et al. [1995].)

For instance, the classifier approach may pick out certain frequency bands that are important in defining certain features. It may also let us explore the best temporal window for finding targets; is it the length of a phone, or is it larger, close to the scope of a syllable?

A classifier will be trained on cases where the phonology is in little doubt and then run on the rest of the data. The classifier will return an estimate of the probability that a particular articulatory target exists at a particular moment; we will then see if the potential target correlates with a certain phone or combination of phones.

3.2 Experiment 2: Real-Time *vs.* Stroboscopic MRI

There are two broad classes of dynamic MRI methods: real-time sequences that collect images from a single production of a sentence and stroboscopic sequences that build up images from repetitive productions over several cycles. If a process is periodic, like a heart beat, the stroboscopic sequences will provide better images with less motion blurring and increased resolution.

Each image in a stroboscopic sequence is assembled from data collected in 10–20 repetitions. We will investigate whether repeating a sentence ten or twenty times gives data that reflect the details of normal, meaningful speech, allowing useful application of these sequences

[1] Not to be confused with a phonological feature. See glossary.

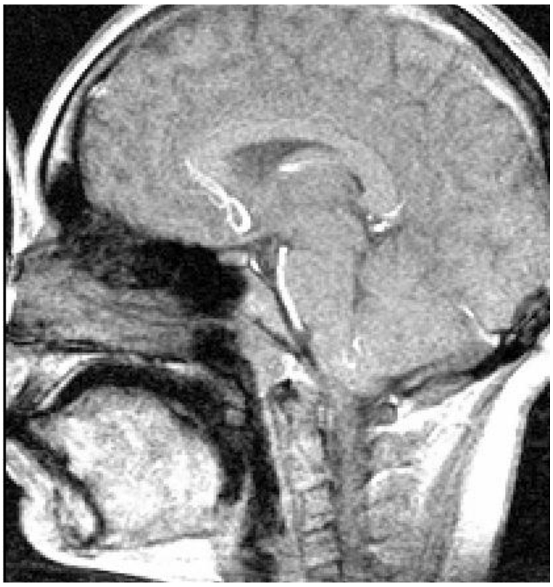


Figure 1: Sample image from a 14 frame-per-second stroboscopic sequence, with 0.9 mm×0.9 mm pixels.

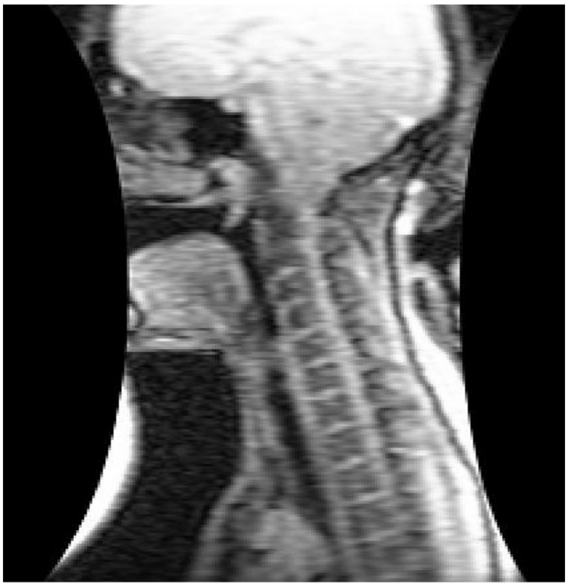


Figure 2: Sample MR image from a 5 frame-per-second real-time movie of speech, with 1.2 mm×1.2 mm pixels.

to speech. (See Figures 1, 2 and `http://www.phon.ox.ac.uk/~jcoleman/Dynamic_MRI.html` for preliminary data).

Timing shifts, inconsistent articulation, or speech errors between the repetitions will degrade the final image and introduce artifacts. We will check for such potential problems by comparing data obtained with stroboscopic sequences to corresponding real-time MR images and by comparing the acoustic properties of different repetitions. We will collect MRI and acoustic data under three production conditions:

1. A list of sentences where each sentence appears once.

2. A list where six instances of each sentence appear randomly (a simulation of the real-time version of Experiment 3).

3. A list of sentences where each is repeated thirteen times in succession. Each repetition is synchronised with a start signal. (In this condition, we will collect MRI data with both stroboscopic and real-time sequences for a direct comparison.)

We will compare the three conditions by perceptual tests, acoustic measurements, and measurements on MR images of the relevant articulators. Images of articulator motions will be collected while the subjects read sentences from a display. We will take mid-sagittal scans to get both velum and tongue data.

The subjects' speech will be recorded through a noise-cancelling microphone. Preliminary experiments show that the audio is easily intelligible. Post-processing the recordings will subtract most of the remaining MRI noise. We expect to reach a +20 dB signal-to-noise ratio, which would permit almost unrestricted analyses.

We will check, first, whether the repetitive speech is within the range of normal variation of read speech and, second, whether we can reliably collect clean images with the stroboscopic sequences. The answers will determine whether we use real-time or stroboscopic sequences for Experiment 3. This work is relevant to many experimental designs in phonetics and phonology and experimental psychology that use repeated stimuli. Experiment 2 will also yield data on the amount of utterance-to-utterance variation for use in the analysis of Experiment 3.

3.3 Experiment 3: Modelling Articulation from Features

In this experiment, we will design a corpus of text to be read, and then measure articulator positions from MRI movies of the resulting speech. From this, we will build models that have a phonological part (a string of features) and a motor-control part that uses those features to predict articulator positions. The motor control strategy will have a number of undetermined parameters that correspond to the articulatory targets for each feature and some parameters that control how the targets interact with each other. How well will it work and what are the limitations?

3.3.1 Design

The sentences for the MRI coarticulation experiment will be selected through Experiment 1 (§3.1) and might be similar to Appendix B. We will use sentences that contain a central VCV section and a controlled context on each side. The central and context sections will be drawn from the same small set of phones. We will control where word and phrase boundaries fall to reduce the effect of prosody on articulation (*c.f.* Fougeron and Keating [1997]). From this data, we can cleanly derive a model of coarticulation because the (knowable) coarticulation from the context section into the centre will normally be stronger than the (unknown) coarticulation from the outskirts.

We will use a set of phones that includes [±ATR] and [±nasal] contrasts, along with phones where these features are unspecified. Initial experiments indicate our newswire corpus contains between 100 and 1000 suitable sequences, with tens of distinct VCV sections in tens of distinct contexts. This gives us a wealth of material from which we can design the corpus. We will edit as necessary.

The real-time imaging option requires six repetitions of each utterance to get an image centred within 25 ms of the middle of the phone under study. The stroboscopic sequence option requires more (thirteen) repetitions, but we anticipate shorter pauses between productions so that the number of distinct utterances will not be too different from the real-time case.

3.3.2 From Images to Articulator Positions

We will measure the tongue and velum shapes and positions from the MR images. We have had preliminary success with an algorithm that amounts to automated replication of manual labelling. The tongue and velum positions are manually marked on about 30 images for each subject. A program then compares new images to the marked set and chooses the library image(s) that are closest to the new data. The tongue and velum shapes will then be estimated from the near-matches. Four masks will focus the comparison onto specific areas, of the tongue and velum. An image that is not a close match will be manually marked then added to the library. The resulting dynamic shapes will be the data for the articulatory model.

3.3.3 Methods of Analysis

In this analysis, we want to learn:

1. how well we can reproduce the data with our phonology/motor control hypothesis,

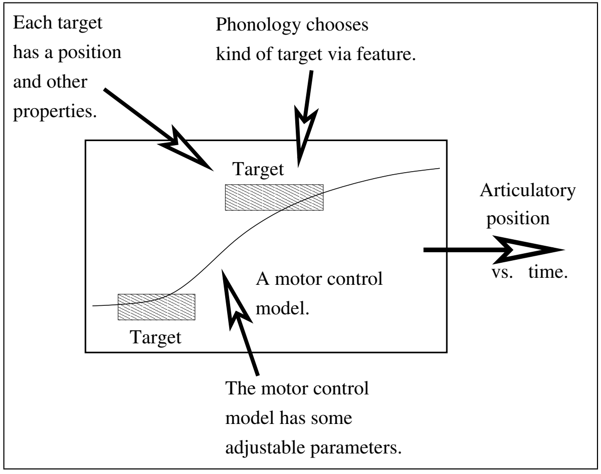
2. what are the properties of the articulatory targets,

3. details of the motor control strategy, and (if necessary)

4. how to adjust the phonological representation to best model the data.

Our model has two steps: first, a phonological step that maps from a sequence of words to relevant features and thence to a sequence of articulatory targets. Each target is centred at a specific time and influences the motion of the tongue or velum around that time. Targets may have different strengths: some may tightly constrain articulator motions, and others may only weakly influence the motions.

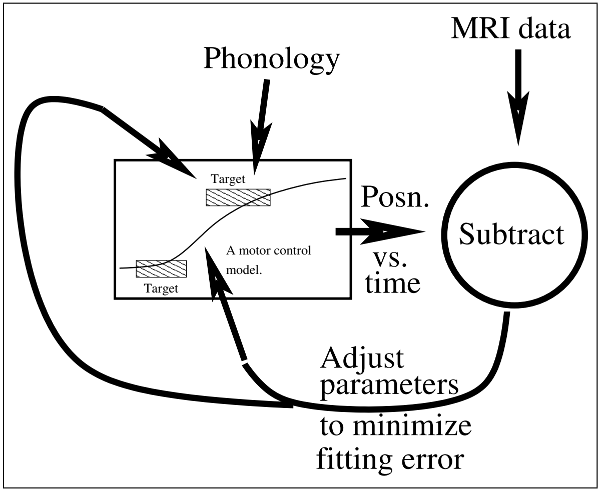
From this sequence of targets, the model's second step then computes a smooth articulatory trajectory that comes near each target (Figure 3.3.3). *Schematic of the motor control model.*


Each target has a position and other properties.
Phonology chooses kind of target via feature.
Target
Articulatory position
vs. time.
A motor control model.
Target
The motor control model has some adjustable parameters.

Different models of motor control correspond to different choices of trajectory. The models of motor control will be supplemented with a definition of distance from each articulatory target [Kochanski and Shih, 2003]. (A task will be to develop software to efficiently compute the articulatory motions for any sequence of targets.)

So far, the motor control models dictate how the targets should interact, and the phonology dictates their approximate placement. However, the timing of targets within their syllable, their spatial position, and how each target interacts with other nearby targets is still undefined.

We will then find optimal values for these target details by fitting models to the corpus and selecting the target details for each relevant feature that best match the data (Figure 3.3.3). *Finding the parameters that give the best fit to MRI data.*



We will also validate the analysis procedure by testing it against the MOCHA corpus [Wrench, 1999]. Specifically, we will check if the low sampling rates of the MRI data (5–15 Hz) may systematically bias the results.

After the optimisation, the error between each model and the corpus then quantifies how good the model is. For instance, we might consider all possible motions for the articulatory target (*e.g.* [+ATR]) and different amounts of strengths for the targets. We will apply the same target with the same properties everywhere that the phonology tells us that [+ATR] should appear, optimise, and measure how well the best model fits.

We will use this strategy to investigate the need for [±ATR] in English. By building several models, we can compare the phonologists' usual proposal of [±back], [±high], [±low] and [±ATR] to alternatives of the same complexity that have more levels of tongue height but not [±ATR] (*e.g.* Harshman et al. [1977]). If one or the other of these models is significantly better at fitting the data, it will tell us whether or not [±ATR] is useful for an articulatorily-based phonological representation.

4 Acknowledgements

Chris Alvey, Bettina Braun, Stephen Golding, Clare Mackay, Burton Rosner, and the University of Oxford Centre for Clinical Magnetic Resonance Research.

5 Word Count

The scientific justification consists of 3498 words, excluding this section.

Appendices

A Glossary

Term	Definition
Advanced Tongue Root (ATR)	A phonological feature that corresponds to the whole of the tongue moving forward, as if pushed from the root.
Articulator	Moveable parts of the body used to produce speech (*e.g.* lips, tongue, jaw, velum, or larynx).
Feature	An abstract object in phonology that is part of the mental representation of a sound. Features are the atomic components of phonology.
Feature vector	(In the field of machine learning.) A list of acoustic measurements that is used to define the important properties of a sound, as input to a classifier.
Invariant	A constant, abstract, underlying property that is shared between a group of diverse surface forms.
Mid-sagittal	The central plane of the body, equidistant from the ears.
Nasal [+nasal]	Indicates that the velum is or should be open.
Phone	A speech sound: a vowel or a consonant.
Phonology	Part of Linguistics, dealing with patterns of contrast between words, in terms of discrete symbols, rather than details of pronunciation.
Read speech, read sentences	Speech produced by reading aloud. In this case, it will be lists of unrelated sentences.
Symbols in []	Phonological features.
Symbols within //	Each character refers to one specific sound.
Tongue root	The posterior aspect of the tongue, near and above the epiglottis.
VCV	Vowel-consonant-vowel sequence.
Velum	The articulator that closes off the nasal cavity. (Soft palate.)
"schwa" (/ə/)	A mid-central vowel that you might find in the word "the" in "Give me the book".
/ɪiʊu/	The vowels in "bit", "beet", "book", "boot".

B Example Sentences

These sentences are selected from a newswire corpus to meet the criteria of both Experiments 1 and 3. They are sequences of seven phones, with a VCV central section, where all vowels are chosen from /ɪ i ʊ u ə/ and consonants from /t s m p/. These sentences were selected to have a word boundary after the third phone. (Word boundaries are shown by spaces.)

Trigram	7-gram	Text
A X B	A X B	
i mi	sti mitɪ	...when scheduling tru**stee meeti**ngs.
i p i	i si p ipə	“W**e see peo**ple laughing who usually don’t laugh”.
i p i	u si p ipə	“**You see peo**ple climbing, and they flow right up the rock”.
i p i	t si p ipə	“We don’**t see peopl**e ordering huge music systems”, Rolf adds.
u mi	ə tu mit ɪ	Th**e two meet i**n a New York park,...
u p i	u tu p ipə	I kn**ew two peopl**e who were almost certainly born slaves, ...
u mɪ	ə tu mɪsɪ	It identified th**e two missi**ng fighter pilots as...
u mɪ	i tu mɪst	...Denver thirty windy; Dallas fift**y two mist**y; Edmonton...
u p ɪ	ə tu p ɪst	...including a version of th**e two pist**ols Denton used.
i mi	ɪti mitɪ	Sixth, require comm**ittee meeti**ngs to be open to the public.
i p i	isi p its	...brings to mind gr**easy pizz**a joints and hookers.
i mɪ	ɪti mɪst	...after the c**ity missed** its third opening date...
i mɪ	ɪti mɪst	That’s pr**etty myst**ical and very powerful.
i p ɪ	pti p ɪst	...the thirty-year-old gunman, still clutching an em**pty pist**ol.
i mə	upi məti	Frances Barber manages to rise above her s**oupy mate**rial.
⋮	⋮	⋮

C Plan of Work

Task / Month of Work	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
• Audio processing algorithms to reduce MRI noise		1	1																								
• Exp1: preparation	J	2																									
• Exp1: data collection			2	2																							
• Exp1: segmentation				c	c	c																					
• Exp1: analysis					2	J	2	2																			
• Exp1: publication									J	2																	
• Exp2: preparation				2																							
• Software to extract articulator positions from MRI images								1		1	1																
• Exp2: data collection											2	2															
• Exp2: segmentation												c	c														
• Exp2: analysis												2	2														
• Exp2: publication														2	2												
• Exp3: implement articulatory models				1	1		1					1	1														
• Exp3: test models on MOCHA														1		1											
• Exp3: preparation															1												
• Exp3: data collection																2	2										
• Exp3: segmentation																	c	c									
• Exp3: from images to articulator positions																	J	2	2								
• Exp3: analysis																		1	1	J	J	1	1				
• Exp3: Publication																								J	J		
• Archive data																							2			1	
• Build web site																						2				2	
• Write project report																											J

- Symbols indicate the primary responsibility: 1=Kochanski, 2=Braun, J=joint, c=contract (segmentation). The contractor will segment the speech data, finding and marking the phones required for analysis. "C" marks the expected duration of the work, which is less than full-time equivalent.
- "Preparation" includes pilot experiments, subject recruitment, and the purchase and installation of experiment-related equipment, and software to collect the data.
- Data collection depends upon the availability of subjects and the MRI system and is thus less than full-time, so we have allowed some overlap with other tasks.
- This work schedule counts working months only, excluding vacation, so the project duration will be 29 months overall.

D MR Scanner Requirements

Experiment 1 will not use MRI. Experiment 2 will require 8 hours of scanner time to cover the five subjects. For Experiment 3, 19 hours of scanner time will be required.

E Ethics Approval

In order to use the available MR scanner on people, full ethics permission must be obtained from the Oxfordshire Regional Ethics Committee. They require that the project must have been externally refereed and all details of funding and procedure must be finalised before application. Therefore we cannot get apply for ethics approval until our application has been refereed and at least conditionally approved.

However, in a recent project [Coleman and Grabe, 2004], we have obtained ethics approval for what is effectively a pilot experiment for this proposal. We can follow the same protocol for subject recruitment and scanning in this project, so we expect that obtaining approval will be straightforward.

These experiments do not involve any unusual experimental or ethical issues, and we will follow standard procedures and guidelines. While the MR imaging involves the use of human subjects, the procedure is non-invasive, commonly used in research, and we have experience with the necessary safety evaluations and informed consent procedures. None of the experiments involve any deception of the subjects, and we will collect no unusually sensitive personal information. Any personal information collected will be anonymized and stored in compliance with the Data Protection Act. We do not expect that the success of the experiment will raise any significant ethical issues.

F Progress Report

Progress of research under ESRC award RES-000-23-0149
to Dr. E. Grabe and Dr. J. S. Coleman,
"A quantitative model of intonational variation in the British Isles"

This project was originally established for the period 01/05/2003 to 30/04/2005. From 01/05/2004, by arrangement with the ESRC, Dr Grabe changed to half-time, and the award was extended to 30/04/2006. The grant application included the salary of a quite junior research assistant for two years, but following the recruitment process we appointed a much more experienced intonation researcher, Dr Gregory Kochanski, to the second position, at a higher salary but for a shorter term (23/03/2003–10/12/2004). The research has progressed very well, and we are on schedule to have completed many of the main milestones by 20/04/2005. The remaining tasks will be completed by Dr Grabe and Dr Coleman by 20/05/2006.

The main questions in the research proposal were:

1. How does the linguistic structure of intonation, dialect, speaking style and context of use, and their interactions, relate to f_0 (the pitch of the voice)?

2. To what extent, and in what ways, do speaker-specific factors and gender differences play a role in English intonation?
3. How can intonation variation be properly represented and modelled, at the linguistic level (*e.g.* variation in intonation transcriptions), at the acoustic level (variation in f_0 contours) and variation in the mapping between these two levels?

We have investigated variation according to dialect, speaking style and sentence type (in the IViE corpus) at the linguistic and acoustic levels. For read sentences, we measured the degree of variation in the intonation contours of different sentence types in three different dialects; a book chapter on this work is in press. We have constructed bi-gram models of intonation patterns, and have determined some differences between the various speech styles and dialects. In particular, Belfast intonation is highly significantly different from the other 6 dialects, even Dublin, because of the very frequent usage of sentence-final rises. Also, there is north-south variation, with Cambridge and London intonation being very similar and Bradford, Leeds and Newcastle forming a second cluster.

Regarding speech styles, read sentences, read passages and dialogues (including free conversation) employ similar intonation patterns. In re-telling a story, however, the patterns are somewhat different, possibly because the additional task of recall from memory in this task.

This work supports our preliminary findings that there are no dialect-specific intonation contours, even though different dialects, styles and sentence types can use the contours in substantially different proportions. It is thus not reasonable to talk about *e.g.* "Newcastle *vs.* Cambridge intonation contours".

Acoustic differences have been examined using more sophisticated methods than were original envisaged in the research proposal. Dr Kochanski has modelled the f_0 contours (and other acoustic parameters, including loudness), using orthogonal polynomials. This curve-fitting technique yields a compact numerical description of each contour.

He then built a classifier for discriminating between sentence types (declarative sentences *vs.* three types of questions) on the basis of their intonation. The classifiers revealed that the differences between sentence types is not just marked at the end of the utterance: these sentence types are distinguished by intonation from the beginning. This work has been disseminated in a conference proceedings paper.

Dr Kochanski has also examined the acoustic differences between prominent and non-prominent syllables, an important aspect of intonation that is an extension of the original proposal. This work has shown that loudness and duration, not f_0, are the main features of prominent syllables in all of the dialects. At least for the speakers in the IViE corpus, the common idea (not well supported experimentally, but in several intonation textbooks) that pitch peaks and motions lend prominence to a syllable seems to be invalid. This work is close to submission as a journal paper.

These results partially answer questions 1 and 3. Question 2 is being attended to at present. We have not yet looked for variation in the mapping between intonation transcriptions and f_0 contours. However, we have developed analytical techniques that can be applied, and we foresee no obstacles to completing those investigations in the remaining period of the award.

G Word Count

The Annexes to the proposal consist of 2891 words, counting the bibliography (below), but not counting this section.

References

A.-P. Benguerel and H. A. Cowan. Coarticulation of upper lip protrusion in French. *Phonetica*, 30:41–55, 1974.

A.-P. Benguerel, H. Hirose, M. Sawashima, and T. Ushijima. Velar coarticulation in French: a fiberscopic study. *J. Phonetics*, 5:149–158, 1977a.

A.-P. Benguerel, H. Hirose, M. Sawashima, and T. Ushijima. Velar coarticulation in French: an electromyographic study. *J. Phonetics*, 5:159–167, 1977b.

C. S. Blackburn and S. Young. A self-learning predictive model of articulator movements during speech production. *J. Acoustical Society of America*, 107(3):1659–1670, March 2000.

C. P. Browman and L. Goldstein. Articulatory gestures as phonological units. *Phonology*, 6:201–251, 1989.

C. P. Browman and L. Goldstein. "Targetless" schwa: an articulatory analysis. In G. J. Docherty and D. R. Ladd, editors, *Papers in Laboratory Phonology*, volume II, pages 26–65. Cambridge University Press, 1992.

C. P. Browman and L. M. Goldstein. Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252, 1986.

G. Brown. Consonant rounding in British English: the status of phonetic descriptions as historical data. In R. E. Asher and E. J. A. Henderson, editors, *Towards a History of Phonetics*, pages 67–76. Edinburgh University Press, 1981.

J. C. Carr, O. Simonetti, J. Bundy, D. Li, S. Pereles, and J. P. Finn. Cine MR angiography of the heart with segmented true fast imaging with steady-state precession. *Radiology*, 219:828–834, 2001.

E. A. Christou, M. Shinohara, and R. M. Enoka. Fluctuations in acceleration during voluntary contractions lead to greater impairment of movement accuracy in old adults. *J. Applied Physiology*, 95(1):373–384, March 2003.

A. C. Cohn. Nasalisation in English. *Phonology*, 10:43–81, 1993.

J. Coleman. Discovering the acoustic correlates of phonological contrasts. *J. Phonetics*, 31: 351–372, 2003.

J. Coleman and E. Grabe. A quantitative model of intonational variation in the British Isles, May 2003. URL `http://www.phon.ox.ac.uk/~esther/oxigen.html`. UK Economic and Social Research Council Award RES-000-23-0149.

J. Coleman and E. Grabe. Larynx movements and intonation in whispered speech, 2004. URL `http://www.phon.ox.ac.uk/~jcoleman/MRI_research_proposal.html`. British Academy Award SG-36269.

D. Demolin, S. Hassid, T. Metens, and A. Soquet. Real-time MRI and articulatory coordination in speech. *Comptes Rendues Biologies*, 325:547–556, 2002.

D. Demolin, V. Lecuit, T. Metens, B. Nazarian, and A. Soquet. Magnetic resonance measurements of the velum port opening. In *Proceedings of the International Conference on Spoken Language Processing*, pages 425–429, 1998. Sydney.

L. Deng. Switching dynamic system models for speech articulation and acoustics. In M. Johnson, M. Ostendorf, S. Khudanpur, and R. Rosenfeld, editors, *Mathematical Foundations of Speech and Language Processing*, pages 115–134. Springer, New York, 2004.

A. G. Feldman. Once more on the equilibrium-point hypothesis (lambda model) for motor control. *J. Motor Behavior*, 18:17–54, 1986.

T. Flash and N. Hogan. The co-ordination of arm movements: An experimentally confirmed mathematical model. *J. Neuroscience*, 5:1688–1703, 1985.

A. K. Foldvik, U. Kristiansen, J. Kværness, and H. de Bonnaventure. A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI). In *Proceedings of Eurospeech '93*, volume 1, pages 557–558, 1993.

C. Fougeron and P. Keating. Articulatory strengthening at the edges of prosodic domains. *J. Acoustical Society of America*, 101:3728–3740, 1997.

C. A. Fowler. Coarticulation and theories of extrinsic timing. *J. Phonetics*, 8:113–133, 1980.

H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage, editor, *The Production of Speech*, pages 39–55. Springer, New York, 1983.

B. Gick. An X-ray investigation of pharyngeal constriction in American English schwa. *Phonetica*, 59:38–48, 2002.

M. Halle and K. Stevens. On the feature Advanced Tongue Root. *Quarterly Progress Reports, MIT Research Lab for Electronics*, 94:209–215, 1969. Republished in *From Memory to Speech and Back*, M. Halle, Mouton de Gruyter, 2003.

C. M. Harris and D. M. Wolpert. Signal-dependent noise determines motor planning. *Nature*, 394:780–784, 1998.

R. Harshman, P. Ladefoged, and L. Goldstein. Factor analysis of tongue shapes. *J. Acoustical Society of America*, 62:693–707, 1977.

S. Hawkins and N. Nguyen. Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *J. Phonetics*, 32:199–231, 2004.

S. Heid and S. Hawkins. An acoustical study of long-domain /r/ and /l/ coarticulation. In *Proceedings of the 5th Seminar on Speech Production: Models and Data*, pages 77–80. ISCA, 2000. Kloster Seeon, Bavaria, Germany.

W. L. Henke. *Dynamic Articulatory Model of Speech Production Using Computer Simulation*. Ph.D., Massachusetts Institute of Technology, Cambridge, MA, 1966.

R. Jakobson, C. G. M. Fant, and M. Halle. *Preliminaries to Speech Analysis: the distinctive features and their correlates*. The MIT Press, Cambridge, MA, 1952.

T. Kaburagi and M. Honda. Dynamic articulatory model based on multidimensional invariant-feature task representation. *J. Acoustical Society America*, 110(1):441–452, July 2001.

P. A. Keating. Underspecification in phonetics. *Phonology*, 5(2):275–292, 1988.

S. King and A. Wrench. Dynamical system modelling of articulator movement. In *Proceedings of the International Conference of the Phonetic Sciences (ICPhS)*, pages 2259–2262, August 1999. San Francisco.

Greg Kochanski and Chilin Shih. Prosody modeling with soft templates. *Speech Communication*, 39(3-4):311–352, February 2003. URL `http://dx.doi.org/10.1016/S0167-6393(02)00047-X`.

Greg Kochanski, Chilin Shih, and Hongyan Jing. Hierarchical structure and word strength prediction of Mandarin prosody. *International J. Speech Technology*, 6(1):33–43, January 2003. ISSN 1381-2416. URL `http://dx.doi.org/10.1023/A:1021095805490`.

E. G. T. Liddell and C. S. Sherrington. Recruitment and some other features of reflex inhibition. *Proceedings of the Royal Society (series B)*, 97:488–518, 1925.

P. Lieberman. Phonetic features and physiology: a reappraisal. *J. Phonetics*, 4:91–112, 1976.

H. Magen. The extent of vowel-to-vowel coarticulation in English. *J. Phonetics*, 25:187–205, 1997.

K. Mathiak, U. Klose, I. Hertrich, W. Grodd, and H. Ackermann. Stroboscopic articulography by fast magnetic resonance imaging. *International J. Communication Disorders*, 35 (419-425), 2000.

S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd. An approach to real-time magnetic resonance imaging for speech production. *J. Acoustical Society of America*, 115(4):1771–177, April 2004.

S. E. G. Öhman. Coarticulation in VCV utterances: Spectrographic measures. *J. Acoustical Society of America*, 39:151–168, 1966.

S. E. G. Öhman. Numerical models of co-articulation. *J. Acoustic Society of America*, 41: 310–320, 1967.

L. P. O'Mard. Development system for auditory modelling. Web page, Centre for the Neural Basis of Hearing, University of Essex, UK, August 2004. URL `http://www.essex.ac.uk/psychology/hearinglab/dsam/index.htm`.

R. D. Patterson, M. H. Allerhand, and C. Giguere. Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform. *J. Acoustical Society of America*, 98:1890–1894, 1995.

J. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, MIT, Cambridge, Massachusetts, 1980.

P. N. Sabes, M. I. Jordan, and D. M. Wolpert. The role of inertial sensitivity in motor planning. *J. Neuroscience*, 18:5948–5957, 1998.

A. Slater and J. Coleman. Non-segmental analysis and synthesis based on a speech database. In H. T. Bunnell and W. Idsardi, editors, *Proceedings of ICSLP 96, Fourth International Conference on Spoken Language Processing*, volume 4, pages 2379–2382, 1996.

P. West. Long-distance coarticulatory effects of British English /l/ and /ɹ/: an EMA, EPG and acoustic study. In *In Procedings of the 5th Speech Production Seminar*, pages 105–108, 2000a. Seeon, Germany.

P. West. Perception of distributed coarticulatory properties of English /l/ and /ɹ/. *J. Phonetics*, 27:405–425, 2000b.

D. H. Whalen. Coarticulation is largely planned. *J. Phonetics*, 18:3–35, 1990.

D. M. Wolpert. Computational approaches to motor control. *Trends in Cognitive Sciences*, 1(6):209–216, September 1997.

A. Wrench. MOCHA-TIMIT. speech database, Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, 1999. URL `http://sls.qmuc.ac.uk`.