BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

AN INVESTIGATION OF ARTICULATORY-ACOUSTIC RELATIONSHIPS IN SPEECH PRODUCTION

by

ALFONSO NIETO-CASTANON

Ingeniería de Telecomunicación, University of Valladolid, 1996

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2004

Approved by

First Reader

Frank H. Guenther, Ph.D.
Associate Professor of Cognitive and Neural Systems

Second Reader

Joseph Perkell, Ph.D.
Adjunct Professor of Cognitive and Neural Systems

Third Reader

Eric Schwartz, Ph.D.
Professor of Cognitive and Neural Systems; Electrical, Computer and Systems Engineering; and Anatomy and Neurobiology

Dedicated to Laura Medrano in exchange for a smile

ACKNOWLEDGMENTS

I would like to acknowledge my network of family, friends, and teachers, who managed to confuse and amaze me by becoming my teachers, family, and friends.

This work has been partially funded by NIH grant R01DC02852 (Frank H. Guenther, P.I.), and FPI grant Fundación Séneca (Juan López Coronado, P.I.)

AN INVESTIGATION OF ARTICULATORY-ACOUSTIC RELATIONSHIPS IN SPEECH PRODUCTION

(Order No.)

ALFONSO NIETO-CASTANON

Boston University Graduate School of Arts and Sciences, 2004

Major Professor: Frank H. Guenther, Associate Professor of Cognitive and Neural Systems

ABSTRACT

This thesis is a combination of empirical and modeling work concerning articulatory-acoustic relationships in speech production. The empirical work investigates the functional relationship between articulatory variability and stability of acoustic cues during American English /r/ production. The analysis of articulatory movements shows that the extent of intra-subject articulatory variability along any given articulatory direction is strongly and inversely related to a measure of acoustic stability (the extent of acoustic variation that displacing the articulators in this direction would produce). The presence and direction of this relationship is consistent with a speech motor control mechanism that uses a third formant frequency target for /r/. Simulations of two speakers' /r/ productions, using the DIVA model of speech production in conjunction with novel speaker-specific vocal tract models derived from magnetic resonance imaging data, mimic the observed range of articulatory gestures each subject used in different phonetic contexts, while exhibiting the same articulatory/acoustic relations as those observed in the experimental data. Overall these results indicate that the production target for American English /r/ is acoustic in nature, rather than articulatory.

Current models of speech production that use acoustic targets drastically simplify the nature and dimensionality of acoustic descriptors in order to facilitate efficient and robust control of the speech apparatus. These simplifications are not in accordance with neurophysiological and imaging evidence on the cortical representation of sounds. The modeling section of this thesis proposes a solution to the speech production control problem that uses a cortical representation of the sound that fits the multifaceted representations found in auditory cortex. A mathematical approximation to the relationship between articulatory movements and the associated changes in the sound spectrum leads to the definition of a novel difference measure for comparing two sound spectra. This measure is closely related to the articulatory movements necessary to transform one sound into the other. A neural model for the cortical representation of sounds is then developed to implement the basic computations leading to estimation of the proposed difference measure. Simulations of this neural model demonstrate successful inverse control strategies for speech production based on acoustic targets.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

DIVA	Directions Into Velocities of Articulators
dof	degrees of freedom
EMMA	Electromagnetic Midsagittal Articulometer
F3	Third formant
FIR	Finite Impulse Response
HRBF	Hyperplane Radial Basis Function
Hz	Hertz
LPC	Linear Predictive Coding
MRI	Magnetic Resonance Imaging
PCA	Principal Component Analysis
TR	Repetition Time.
VCV	Vowel-Consonant-Vowel

CHAPTER 1. A MODELING INVESTIGATION OF ARTICULATORY VARIABILITY AND ACOUSTIC STABILITY DURING AMERICAN ENGLISH /R/ PRODUCTION

I. INTRODUCTION

When producing a given phoneme, speakers use a set of articulators (e.g. tongue, jaw, lips) to affect the vocal tract shape and, ultimately, the characteristics of the resulting acoustic signal. The vocal tract configuration for the production of a given phoneme is not uniquely defined by phoneme identity. Different speakers will use different articulatory configurations when producing the same phoneme, and often the same speaker will use a range of different articulatory configurations when producing the same phoneme in different contexts. In particular, the American English phoneme /r/ has been associated with a large amount of articulatory variability (Delattre and Freeman, 1968; Espy-Wilson and Boyce, 1994; Guenther et al., 1999). While large, the degree of articulatory variability present in natural speech does not seem to hinder phoneme recognition by listeners, and it is often conceptualized as an expression of control mechanisms that make efficient use of a redundant articulatory system. Such efficient use of redundancy in biological motor systems is often referred to as *motor equivalence*.

Current speech movement control theories dealing with the motor equivalence problem can be roughly classified depending on the type of phonemic targets hypothesized (see MacNeilage, 1970, for motivations of a target-based approach to speech motor control theories). The task-dynamic model of Saltzman and Munhall (1989) exemplifies a type of computational model in which phonemic targets are characterized in terms of *tract variables* representing specific aspects of the vocal tract shape that can be independently controlled by the speech control mechanism

(e.g., lip aperture, tongue dorsum constriction location, etc.) In this model, articulatory variability can arise as a consequence of "blending" effects from the context phonemes. For example, when producing a /b/ in a VCV context, a full bilabial closure represents the targeted tract variable. Other aspects of the vocal tract not affecting the targeted tract variable, such as tongue shape, will vary depending on the shape adopted in the production of the leading vowel, while also being subject to anticipatory movements towards the following vowel configuration. In this way, articulatory variability in different phonetic contexts would reflect the interplay between constraints imposed by current and contextual phonemic targets.

The DIVA model (e.g., Guenther et al., 1998; Guenther et al., 2003) exemplifies a second type of computational model of speech motor control in which the phonemic targets are characterized in terms of *acoustic/auditory variables*[1] (for example, formant frequency descriptors). In this model, the control mechanism moves the articulators in the direction that would bring the formants of the resulting auditory signal closest to the targeted formants, without reference to an explicit vocal tract shape target. Articulatory variability then arises naturally as a consequence of the many-to-one mapping between the articulatory configurations and the audible acoustic characteristics of the produced sound. In other words, for these models articulatory variability reflects the variety of articulatory configurations that would produce the desired acoustic properties.

[1] The current version of the DIVA model (Guenther et al., 2003) uses a combination of auditory and somatosensory targets. As a result of learning in the model, sounds whose characteristic acoustic signal can be produced with a wide range of articulator shapes end up with primarily auditory targets, while sounds that can only be produced with a consistent somatosensory pattern (e.g., lip tactile information signaling full closure for a bilabial stop) will have both auditory and somatosensory targets. In other words, the model hypothesizes that the exact nature of the target (auditory and/or somatosensory) for a sound will depend on the amount of variability in the two spaces that is allowable for that sound in the infant's native language. In this chapter we will deal only with /r/, which we believe to have a primarily auditory target in American English.

Often (e.g. Saltzman and Munhall, 1989; Guenther, 1998) the distinction is emphasized between the articulatory configurations (the state of articulatory variables, such as jaw aperture) and the resulting vocal tract shapes (the state of tract variables, such as tongue dorsum constriction degree). This highlights the redundancy of the speech articulatory system. For example, a particular tongue dorsum constriction degree can be achieved with a relatively low jaw height and a relatively high tongue body height (relative to the jaw) or a higher jaw height and lower tongue body height. More generally, both articulatory and tract variables represent different coordinate frames that can be used to represent the state of the vocal tract apparatus (see MacNeilage, 1970, for an introduction on the concept of coordinate systems in speech production). Tract variables represent a more abstract coordinate frame than articulatory variables, since there is a one-to-many relation between tract variables and articulatory variables defined by the geometrical relations among them. In the same way, acoustic or auditory variables (Guenther et al., 1998) can be simply thought of as yet another coordinate frame for the representation of the articulatory state. They represent a more abstract coordinate frame than either tract or articulatory variables, in that there are one-to-many relations between auditory and tract variables, and between tract and articulatory variables. The analysis of variability in articulatory configurations in the production of a given phoneme, similarly to the analysis of errors in a pointing task (Carozzo et al., 1999; McIntyre et al., 2000), is a useful approach for uncovering an appropriate coordinate-frame for the representation of targets in speech production. We thus believe that the analysis of articulatory variability should serve to direct the definition of motor control models of speech production. Based on this view, the goal of the current work is two-fold: 1) to characterize, in a paradigmatic example of articulatory variability (American English /r/), the extent of articulatory variability in relation to hypothesized target representations (relevant tract and acoustic variables); and 2) to test whether a model of speech motor control

based on an acoustic target definition, together with a speaker-specific vocal tract model, can explain the specificities of the observed articulatory variability in individual speakers. To these ends, we first present new, model-based analyses of electromagnetic midsagittal articulometer (EMMA) data on seven subjects from a previous study (Guenther et al., 1999). These analyses characterize the experimentally observed articulatory variability in relation to hypothesized target variables. We then provide simulation results of an auditory target model controlling the movements of speaker-specific vocal tract models based on magnetic resonance imaging (MRI) scans of the vocal tracts of two of the seven experimental subjects. The model movements are then compared to those of the modeled speakers. Note that the present study addresses only the production of American English /r/. Several aspects of this chapter's methodology (to be described later) are specific to the class of vowel and semivowel productions. The extent to which the presented results generalize to the production of other phoneme classes (in particular, consonants) can only be addressed by further studies.

II. METHODS

A. EMMA data collection and analysis

An EMMA system (Perkell et al, 1992) was used to track the movement of six transducer coils on the tongue, jaw, and lips in the midsagittal plane during the production of /r/ in five different phonetic contexts ("warav", "wabrav", "wavrav", "wagrav", "wadrav") for seven American English speakers. A directional microphone was used to record the subjects' speech simultaneously with the EMMA signals. The details of the methodology are described in Guenther et al. (1999). The primary acoustic cue for /r/ is a deep dip in the trajectory of the third formant frequency, or F3 (Boyce and Espy-Wilson, 1997; Delattre and Freeman, 1968). The acoustic signal was therefore processed to extract the F3 trajectory. A linear fit was used to estimate a first-order approximation of the effect of each transducer position on F3 for each subject. The articulatory data were analyzed in terms of the articulatory covariance of the transducer positions at the time of the F3 minimum for /r/ (the acoustic "center" of the /r/) across the different productions.

Previous analyses (Guenther et al. 1999) had shown that articulatory tradeoffs during /r/ production acted to reduce F3 variability. In this section we attempt to assess this kind of finding in the context of different speech motor control theories by testing the ability of theoretically motivated phonemic target variables to predict the observed variability in articulatory configurations. Our rationale is exemplified in Figure 1.1. Let us only consider the movement of the tongue tip in this example. Imagine, during the production of a hypothetical phoneme, the phonemic target consists of accomplishing a given tongue tip constriction degree (distance between the tongue tip and the hard palate). The expected array of final configurations of the tongue tip for the production of this phoneme would be expected to take the approximate form shown in Figure 1.1 left. The axes labeled A and B represent the directions of articulatory

movement resulting from a principal component analysis (PCA) of the final articulatory covariance of a number of productions of the phoneme, and the gray arrow characterizes the direction of articulatory movement affecting the degree of the tongue tip constriction the most. The right side of Figure 1.1 plots for each articulatory direction (A and B) their effect on the hypothesized target variable (*effect on constriction degree*) on the x axis, and their extent of *articulatory variability* on the y axis. This plots schematizes the observation that those articulatory dimensions affecting the target variable the most (B, in this case) would be expected to show a lesser extent of articulatory variability than those dimensions affecting the target variable the least (A). Such differences in the extent of articulatory variability have been, for example, found in the production of the vowels /i/ and /a/, with minimal variability along acoustically critical directions perpendicular to the vocal-tract midline (Perkell and Nelson, 1985). The analyses in this section extend the simple scheme exemplified in Figure 1.1 (with only one transducer reflecting the tongue tip position) to the case of multiple transducers (6 transducers, reflecting tongue, jaw, and lips configurations). The simultaneous analysis of multiple transducers on different articulators allows the articulatory dimensions (12 for each subject) that result from a principal components analysis to characterize complex movements of one or several articulators, such as those described in the literature as trading relations between and within articulators (for example a simultaneous raising of the tongue back and decrease of lip rounding for the vowel /u/, as in Perkell et al, 1995; or a simultaneous raising of the tongue tip and lowering of the tongue back for /r/ as in Guenther et al, 1999). As in the example shown here, a functional relationship between the extent of articulatory variability along each of the resulting articulatory dimensions and their associated effect on a hypothesized target variable is taken as indicative of the use of a specific target scheme in the articulatory movement data being analyzed.

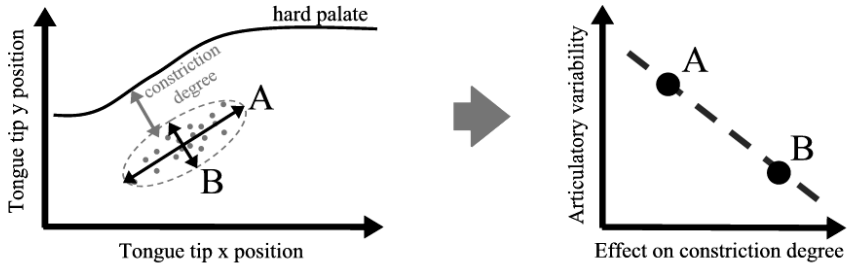

Figure 1.1. Schematic example of articulatory variability analysis for a single articulatory measure of interest (tongue tip position). **Left:** Hypothetical configuration of tongue tip positions in the production of a phoneme that could be characterized by a tongue tip constriction degree phonemic target (anterior is toward the left). A and B represent the directions of the tongue tip movement resulting from a principal component analysis (PCA) of the tongue tip articulatory covariance of multiple repetitions. The arrow labeled constriction degree represents the direction of the tongue tip movement affecting the constriction-degree the most. **Right:** Plot relating the extent of articulatory variability along each of the articulatory directions (A and B) versus the effect that each of these directions has on the hypothetical target variable (constriction degree). Evidence for a tongue tip constriction-degree phonemic target comes from the relative separability of the articulatory variability based on the effect that each of these directions has on the tongue tip constriction degree; the articulatory direction A affects the constriction degree the least and shows a larger extent of articulatory variability, while the articulatory direction B affects the constriction degree the most and shows a smaller extent of articulatory variability. The actual analyses performed in this section attempt to provide evidence for several theoretically motivated phonemic target definitions by extending this simple scheme to the case of multiple articulatory measures of interest (indicated by six transducer positions located on the tongue, lips, and jaw of the speakers).

For each subject, a principal component analysis (PCA) of the articulatory covariance led to the definition of a set of 12 vectors (principal articulatory directions) defining a base in the

articulatory space. Plots were constructed relating, for each articulatory direction, the observed articulatory variability along this direction and: a) its effect on the acoustic cue F3 (as estimated by the original linear fit, in Hertz/mm); and b) its effect on each of the tongue shape indicators (associated with the tongue tip, tongue dorsum, and tongue back constriction location and degree tract variables) as estimated by the principal articulatory direction loadings on the corresponding transducer indicators (in mm/mm units). The directions of movement of the three tongue transducers perpendicular to the palate surface for each subject were used as approximate indicators of the tongue tip, tongue dorsum, and tongue back constriction degree, respectively. Constriction location was defined as the direction perpendicular to the constriction degree (parallel to the palate surface). Data from different subjects were overlaid after normalizing by each subject's total amount of articulatory variability (each subject's total variability was arbitrarily scaled to 10mm). The resulting plots were fit using a linear regression on the log variables. R^2 and p values, as well as confidence intervals for the linear fit parameters, are reported in the Results.

B. Construction of speaker-specific vocal tract models

A *speaker-specific vocal tract model* is a characterization of the range of configurations a speaker's vocal tract could adopt, together with the acoustic output any configuration would produce under glottal excitation. To estimate the former, a set of 2-D MRI midsagittal profiles was acquired for two subjects (the first two subjects in the EMMA experiment) while producing a set of phonemes. To estimate the latter (the associated acoustic outputs), acoustic data were collected at the start of each scan. Data for 27 and 15 phoneme productions were acquired for subject 1 and 2, respectively. The following paragraphs describe the procedure used to interpolate and generalize from the limited available articulatory and acoustic data to other non-observed

configurations. The results provide a simple characterization of the full range of articulatory configurations and acoustic outputs a speaker can produce.

Analysis of vocal tract configurations. Previous approaches to the creation of a parametric description of articulatory movements (e.g. Perrier et al., 1992; Story et al., 1996, 1998) create a grid in the midsagittal plane and obtain the vocal tract area function from the intersection of this grid with the vocal tract outline. An articulatory model based directly on a vocal tract area function representation is, nevertheless, unlikely to produce optimally realistic articulatory movements, given the discontinuity between natural vocal tract articulator movements and the corresponding area function representation using the grid method. In this work we chose to create a parametric definition of the articulator space from a principal component decomposition of the outlines of different vocal tract segments (tongue, jaw and lips). In this way the resulting characterization is expected to be both articulatorily meaningful and continuous with respect to movement of the articulators. T1 (14s TR, 4mm slice) midsagittal MR images were obtained while the subject was producing a set of simple phonetic utterances involving a variety of vowels, semivowels, fricatives, and stop consonants chosen to represent the full range of articulations. Each subject was asked to produce a simple utterance (either a steady-state vowel or a /VC/ sequence) and hold the last phoneme during the 14 seconds of the image acquisition procedure. Images were inspected visually for movement artifacts, and trials with a large amount of movement were removed from further analyses. In each resulting raw MR image, the intensity histogram was clustered into eight representative levels and the air regions were identified using a flood fill algorithm starting from a user-defined point. The resulting air regions were corrected manually to form a connected set defining the air-cavity region of the vocal tract and its outline was extracted for each image. Figure 1.2 shows an example of the estimated vocal tract outline (thick line) and the original MRI data for the first subject. The resulting vocal tract outlines were

aligned spatially using the hard palate outline to correct for subject movement in the scanner. They were then parcellated into different segments of interest (jaw, lips, hard palate, velum, laryngeal region). The tongue body segment is highlighted in Figure 1.2. Each segment was interpolated by a fixed number of equally spaced 2-d points along the identified segment outline in order to obtain a vector descriptor of its shape. For the present study we concentrated on the effect of tongue, lower lips and jaw. Principal component analysis was applied to each of these shape descriptors to obtain a set of five articulatory components: three for the tongue body, and one each for the jaw and lower lip. The variability in articulatory configurations explained by movements of the jaw was removed prior to the estimation of the tongue and lip principal components in order to remove redundancies in their definition (c.f. Maeda, 1990). The resulting set of principal articulatory components was used as a characterization of the range of articulatory configurations the subject could produce. In this way, any articulatory configuration the subject's vocal tract model could produce was represented by a five-element vector, describing the contribution of each of the five articulatory components to the vocal tract shape.

Analysis of acoustic signals and the articulatory to acoustic mapping. Acoustic recordings of the subject's production of each utterance made while in the MRI scanner (just before the onset of the scanner noise) were analyzed using Linear Predictive Coding (LPC) (p=26, Fs=22KHz). The acoustic signal was pre-emphasized with a single delay FIR filter (a1=.95) to reduce the effects due to radiation and the glottal pulse (Wakita, 1973). The first three formant values were extracted for each production.

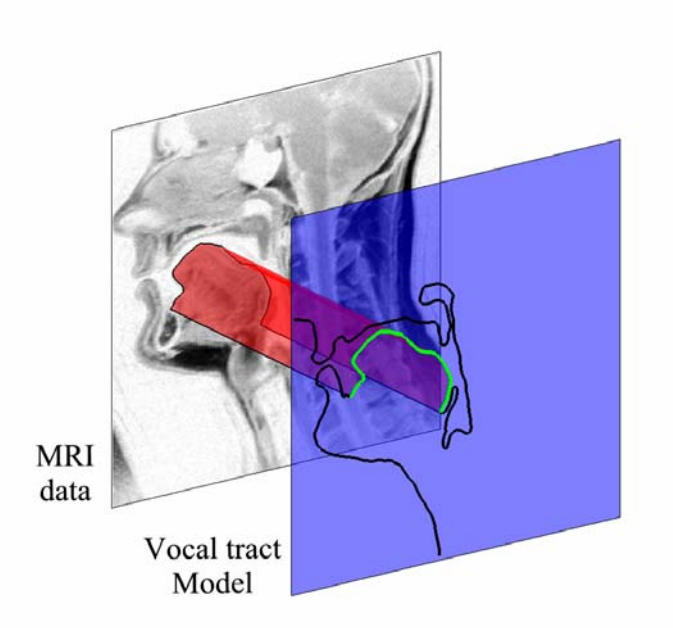

Figure 1.2. Example of MRI midsagittal data and the estimated vocal tract outline (thick line) for Subject 1during the production of /r/. The vocal tract outline was parcellated into relevant segments of interest. The segment corresponding to the tongue shape is highlighted in this plot. MRI images were acquired for each subject while producing a set of simple sustained phonetic utterances (including vowels, semivowels, fricatives, and stop consonants). These data together with the corresponding acoustic signal for each utterance were used to create speaker-specific vocal tract models for two subjects.

In order to approximate the vocal-tract articulatory/acoustic mapping, previous models have derived it from the vocal tract area function either using an elliptical approximation that is difficult to tune (Maeda, 1990) or an elegant but more complex estimation procedure based on multiple 3-D volumetric MRI representations of the vocal tract (e.g. Tiede and Yehia, 1996). The collection of 3-D volumetric data for multiple phonemes is time-consuming and can suffer from problems in determining the location of the teeth, which do not show up on MR images and thus adversely affect the measured area function. Compared to this approach, in this work we used a simple linear mapping to fit the relationship between the articulatory and formant descriptors for

each subject. In this way, the proposed model requires a relatively small amount of MRI and acoustic data for each subject and avoids the complications derived from the estimation of the area function. The validity of this approach was first estimated by creating a random sample of vocal tract configurations, and computing the corresponding acoustic outputs using a standard articulatory synthesizer (Maeda, 1990). A random set of 10,000 valid articulatory configurations was created using a normal distribution of the model's articulatory parameters (mean zero, standard deviation one) hard-limiting between –3 to 3 standard deviations (the full valid range of articulatory parameters in Maeda's 1990 vocal tract model). For this data we found a very significant linear relationship ($R^2 = 0.97$) between the articulatory and formant descriptors. Deviations from linearity were most apparent in extreme configurations (close to a closure). This result indicates that a linear mapping between articulatory and acoustic dimensions is reasonable if the vocal tract is restricted to non-extreme configurations (e.g. vowels and semivowels). In other words, this methodology would not be appropriate for modeling many consonant productions. The linear mapping best fitting the relationship between articulatory and acoustic components for each subject's data was then estimated using linear regression on the articulatory and acoustic descriptors from vowels and semivowels (9 and 6 configurations for Subjects 1 and 2, respectively).

C. Simulations of /r/ production

The DIVA model (Guenther et al., 1998) was used as a controller for the movement of the speaker-specific vocal tract articulators to produce an acoustic /r/ target in different phonetic contexts. The DIVA model can be characterized as a derivative controller in acoustic space. The implementation reduces, at each time-point, to iteratively moving the articulators in the articulatory direction that brings the current acoustic output closest to the desired acoustic target. In mathematical terms, the model uses a pseudoinverse of the Jacobian matrix relating articulator

movements to their acoustic consequences to move in a straight line (in acoustic space) to the target (see Guenther et al., 1998 for details). While in the complete DIVA model this is accomplished by learning this pseudoinverse transformation through experience (e.g. Guenther et al., 1998), in the current implementation we used an explicit calculation of the pseudo-inverse of the articulator-to-acoustic mapping. The articulatory space was defined in terms of the PCA components as described above, and the acoustic space was defined in terms of the first three formants of the spectrum (in Hz). The acoustic target in the model was defined from each subject's own /r/ production formants. To compare the results of the DIVA model simulations to the experimentally obtained EMMA data for each subject, the estimated transducer locations were manually identified on a rest configuration of the modeled speaker-specific vocal tract. The approximate location where the tongue transducers were placed was visually identified following the directives of the original EMMA experimental paradigm, as 1, 2.5, and 5 cm back from the tongue tip. The initial vocal tract configurations of three phonetic contexts (/ar/, /dr/, and /gr/) were manually edited from the original MRI data to approximate the observed initial transducer configuration (75 ms before F3 minimum) in the corresponding contexts for each subject. Simulations of the DIVA model were run starting from these configurations to a "final" configuration at the F3 minimum for /r/. The estimated direction of movement (difference between the final and starting transducer positions) was compared to the measured transducer movement in the same contexts (correlation coefficients are reported). Finally, using all available MRI configurations as initial vocal tract configurations (not just the three used for the preceding analyses), additional simulations were run using the same acoustic target for /r/, and the resulting articulatory variability across the model's /r/ productions was determined. On this data we performed similar articulatory variability analyses as those performed on the original EMMA data.

III. RESULTS

A. Predictive relations between hypothetical target variables and articulatory variability

This section deals with the analysis of articulatory movement data in an attempt to show the ability of different phonemic target hypotheses to account for the observed articulatory variability in the production of /r/. In particular, it was expected that the choice of an "appropriate" phonemic target would provide good separability of those directions of articulatory movement showing large versus small articulatory variability. The main result shows that for any direction of articulatory movement, its effect on an acoustic variable (F3, corresponding to an acoustic phonemic target representation hypothesis) is a good predictive measure of the extent of articulatory variability along this direction of articulatory movement. On the other hand, none of the articulatory variables tested (corresponding to an articulatory phonemic target representation hypothesis) provide as good predictability of the observed articulatory variability.

A first-order approximation to the effect of each of the EMMA transducer positions on F3 was first estimated using linear regression for each subject ($p \leq 3*10^{-6}$; average $R^2=.56$; dof$\geq$71). The EMMA transducer positions at the F3 minimum for /r/ for each of the seven subjects were then analyzed (principal component analysis) to obtain a set of 12 articulatory dimensions characterizing each subject's articulatory covariance (see Section II.A for details). These articulatory dimensions represent deviations of the articulators around an average /r/ configuration for each subject. For a given subject, some of these articulatory dimensions show a relatively high degree of variability (meaning that this subject's /r/ configurations differed relatively largely along those articulatory dimensions), and others show a relatively small degree of variability (meaning that the subject's final articulatory configurations would be relatively stable along those articulatory dimensions). Similarly, each articulatory dimension would also vary with respect to its effect on F3 (meaning that moving the articulators along this direction

produces a relatively large/small change on the F3 acoustic cue) and with respect to its effect on different tract variables (e.g. a large tongue dorsum constriction degree effect means that moving the articulators along this direction produces a large change in the size of the palatal - tongue dorsum constriction). Figure 1.3 shows log plots of the relation between the extent of articulatory variability (vertical axis) and (from left to right) the effect on acoustic (F3) and tract (tongue dorsum degree, tongue tip constriction degree and location) variables. The results show that the degree of articulatory variability along any articulatory dimension is strongly related to its effect on F3 ($R^2 = .46$), but it is not related to any of the tested tract variables ($R^2 \leq .02$). The direction of the observed relation between articulatory variability and effect on F3 is consistent with that expected from a control mechanism using an F3 target; i.e. the final articulatory variability is lower for those articulatory directions most relevant to determining the F3 value. The tract variables chosen in these plots reflect those that would seem most relevant for /r/ production (see Guenther et al., 1999). Correlations with the remaining tongue constriction indicators (tongue dorsum constriction location, and tongue back constriction width and location) were also low ($R^2 \leq .04$).

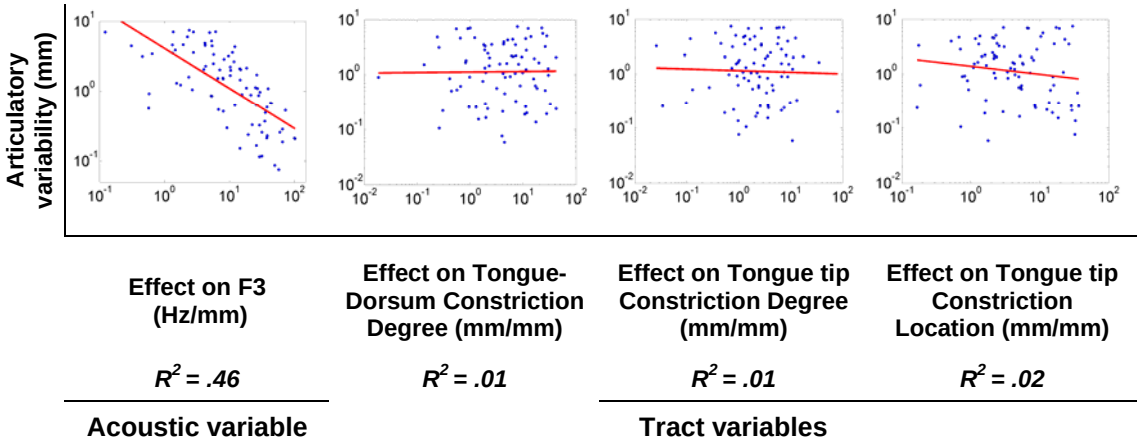

Figure 1.3. Relation of hypothetical target variables to articulatory variability during /r/ production. Each point in the plots represents an articulatory dimension (a direction of movement of the articulators) for a given subject. Plots show the experimentally measured relation between each articulatory dimension's variability and (from left to right) its effect on F3, its effect on tongue-dorsum constriction degree, its effect on tongue tip constriction degree, and its effect on tongue tip constriction location. The strong relation (R^2 = .46) found between an acoustically relevant variable (F3) and the articulatory variability in /r/ production is consistent with that expected from an acoustic target speech motor control mechanism (i.e. the final articulatory variability is lower for those articulatory dimensions most relevant to determining the F3 value). No evidence ($R^2 \leq .02$) of a vocal tract shape- target speech motor control mechanism was found (i.e. the final articulatory variability was not consistently lower for those articulatory dimensions most relevant to determining hypothetical tract variable targets).

The main positive result of the previous analyses is the following: *the extent to which each articulatory dimension was found to vary showed a strong inverse relation to the change in F3 associated with moving the articulators along that same articulatory dimension*. In particular, the standard deviation of the articulatory variability along any given articulatory dimension was well

approximated by the squared inverse of the change (in Hertz) in F3 associated with that articulatory dimension. An F-test on the global fit (linear regression on log variables) for the combined results of all subjects showed a strong association between these variables ($F_{1,82}$ = 69.68; $p = 1 * 10^{-12}$). This fit is shown in the top panel of Figure 1.4.

These results indicate that if, for a given subject, deviating from an average /r/ configuration along a given articulatory direction was found to have a relatively large impact on F3 (low F3 stability), then that subject tended to show little articulatory variability along this articulatory dimension. Conversely, if deviating along a given articulatory direction was found to have relatively little impact on F3 (high F3 stability), then the subject tended to show a larger amount of articulatory variability along this articulatory dimension. We will refer to this as a ***predictive relationship between acoustic stability and articulatory variability***.

To assess the statistical significance of the experimental results we performed a Monte Carlo test (replication of all the analysis steps using a series of simulated datasets conforming to a pre-defined null hypothesis). The null-hypothesis represents the case where there is no relation between articulatory variability and acoustic stability. It is important to note that this null-hypothesis would not be appropriately tested using the significance level of the articulatory/acoustic linear regression R^2's in Figure 1.3. This would provide an increased false-positive rate because such relations can appear solely from noise in the F3 measurement. This is worth clarifying, not only to reassure the reader that the observed relations are not artifactual, but moreover because we feel it helps better understand the nature of the measured articulatory/acoustic relationship. The observed predictive relations indicate that the effect on F3 (in Hz/mm) relates inversely to the articulatory variability (in mm). This is equivalent to stating that the change in F3 (in Hz) is roughly constant when the articulators move one standard deviation (dimensionless, measured in degrees of articulatory variability) along any given

articulatory dimension, no matter if this movement represents a very large or very small change in articulatory configurations. This exactly is the hallmark of a controller using an F3 target: it would distribute the F3 error equally along all available articulatory dimensions, making the "important" articulatory dimensions vary little while the "unimportant" ones vary a lot. However, if the measurement noise in F3 was large (making the measured F3 independent of the articulatory state), artifactual articulatory/acoustic relations would appear in the data. The reason is that in this case the F3 variability (in Hz) would also be found to be independent of the degree of articulatory movement (in standard deviations). This artifactual case can nevertheless be discerned from a non-artifactual scenario based on the degree of association between the articulatory configurations and the acoustic variable F3 (this would be minimal in the case of a noisy F3 measurement). We performed Monte Carlo tests constrained to the observed level of association between these variables to evaluate the possibility of measurement noise inducing the observed articulatory/acoustic relations. A Monte Carlo test provides appropriate false-positive control by directly estimating the distribution of expected R^2 values under the null-hypothesis. 1000 simulated datasets representing the null-hypothesis were constructed. In each of them a simulated target variable was defined by adding Gaussian noise to a tract variable. The amount of noise in the simulated target variable was fixed such that the transducer positions / simulated target variable mapping would have the same association strength as the measured transducer positions / F3 mapping. For each of these datasets we estimated the relation between articulatory variability and the effect on the simulated target variable, and constructed the distribution of expected R^2 values under the null hypothesis. The results show that while the average relation strength expected under the null hypothesis was relatively large (R^2=.29), the probability of obtaining a value as high as the experimentally observed articulatory/acoustic relation (R^2=.46)

was significantly low ($p=4*10^{-3}$). These results indicate that the observed relation is statistically significant beyond possible artifactual causes.

To explore the consistency of the observed articulatory/acoustic relation across subjects, we divided each subject's articulatory dimensions into two equal-size sets: one with the one half of the articulatory dimensions that were associated with lower rates of F3 change (labeled "Small effect on F3 components"), and the other with those articulatory dimensions associated with higher rates of F3 change (labeled "Large effect on F3 components"). The small effect on F3 components were shown to consistently (for all subjects) explain a markedly larger percentage of the total articulatory variability (ranging from 84% to 98%, average 93%) than the large effect on F3 components (2% to 16%, average 7%). These results are shown in Figure 1.4 bottom left. Despite the relatively low number of components (12) available for each subject, a Monte Carlo test (random set definition on the estimated variance components) shows these results to be significant ($p<.05$) for 6 out of 7 subjects (all but Subject 7, for which $p=.08$). Overall, these results indicate that the articulatory variability for each subject shows a strong deviation towards those articulatory dimensions that have a small effect on F3.

To explore the possibility that predictive relations using tract variables (those shown in Figure 1.3 right) failed to become apparent because each subject might use a different articulatory strategy (subject-dependent tract variables) we repeated the across-subject analyses explained above using the tract variables (as opposed to the effect on F3) to create the different large-effect vs. small-effect sets of articulatory components for each subject. We then selected for each subject the tract variable (among the six tongue tract variables) that related best to the articulatory variability. The percentage of articulatory variability explained by the small-effect articulatory components on each subject's optimal tract variable ranged from 56% (Subject 4) to 91% (Subject 6), with an average of 67%. These results are still not as good ($p=.02$; Wilcoxon sign

rank test) as those obtained from an hypothesized acoustic target representation (small-effect components explaining between 84% to 98%, average 93%, of the articulatory variability).

An important source of contextual variability in the current experimental setup is the phonetic context preceding the /r/ production. Articulatory target models often employ context-dependent articulatory targets (e.g., blended targets in the task-dynamic model of Saltzman and Munhall, 1989). The following analyses address two questions regarding the influence of phonetic context: a) whether the observed articulatory/acoustic relations were the result of a common control mechanism or a context-dependent target definition; and b) whether predictive relations using tract variables would become apparent when context-dependent articulatory targets are employed. For these analyses we broke down the total articulatory variability into two components: intra- and inter-context variability. The inter-context articulatory variability is the variability related to the identity of the preceding phoneme. To compute inter-context variability, we first average the /r/ configurations within a given phonetic context, then compute the variability of these averaged configurations. The intra-context articulatory variability is the remaining portion of the articulatory variability, which reflects the range of /r/ configurations for different productions in the same leading phonetic context (i.e. the variability of /r/ configurations in an /ar/ context, averaged with the variability of /r/ configurations in a /dr/ context, etc.) If the observed articulatory/acoustic relations were the result of a context-dependent target definition, this relation would be apparent when looking at the inter-context variability, but not when looking at the intra-context variability. The results (see Figure 1.4 bottom right), on the other hand, show that the articulatory/acoustic relations appear for both the intra-context (R^2=.46) as well as inter-context (R^2=.62) variability, pointing to the action of a common control mechanism across speakers and contexts, rather than a context-dependent target definition.

To explore the second question, i.e. whether predictive relations between tract variables and articulatory variability would become apparent when introducing appropriate contextual information, we repeated the previous context-dependent analyses using hypothesized tract variable instead of acoustic representations. The relations between effect on each of the tested tract variables and intra-context articulatory variability were weak (with R^2 values always lower than .10). These results indicate that using context-dependent articulatory targets does not significantly improve the ability of tract variable target representations to explain the observed articulatory variability. For completeness the relation between the hypothesized tract variables and inter-context articulatory variability was computed and found to be similarly weak (R^2<.02).

Overall, the positive results in this section highlight a strong and consistent relationship between acoustic variables and articulatory variability. This result is schematized in Figure 1.5. This relationship is consistent with that expected from a control mechanism using an F3 target (i.e. the final articulatory variability is lower for those articulatory directions most relevant to determining the F3 value). Furthermore, this relationship appears both when looking at the total articulatory variability and when looking at the intra-context articulatory variability (the articulatory variability in each of the phonetic contexts tested). These results suggest that an acoustic target motor control mechanism utilizing the same acoustic target across contexts can account for the observed range of articulatory configurations during /r/ production. The next sub-section further investigates this assertion with a specific model utilizing an acoustic target for /r/.

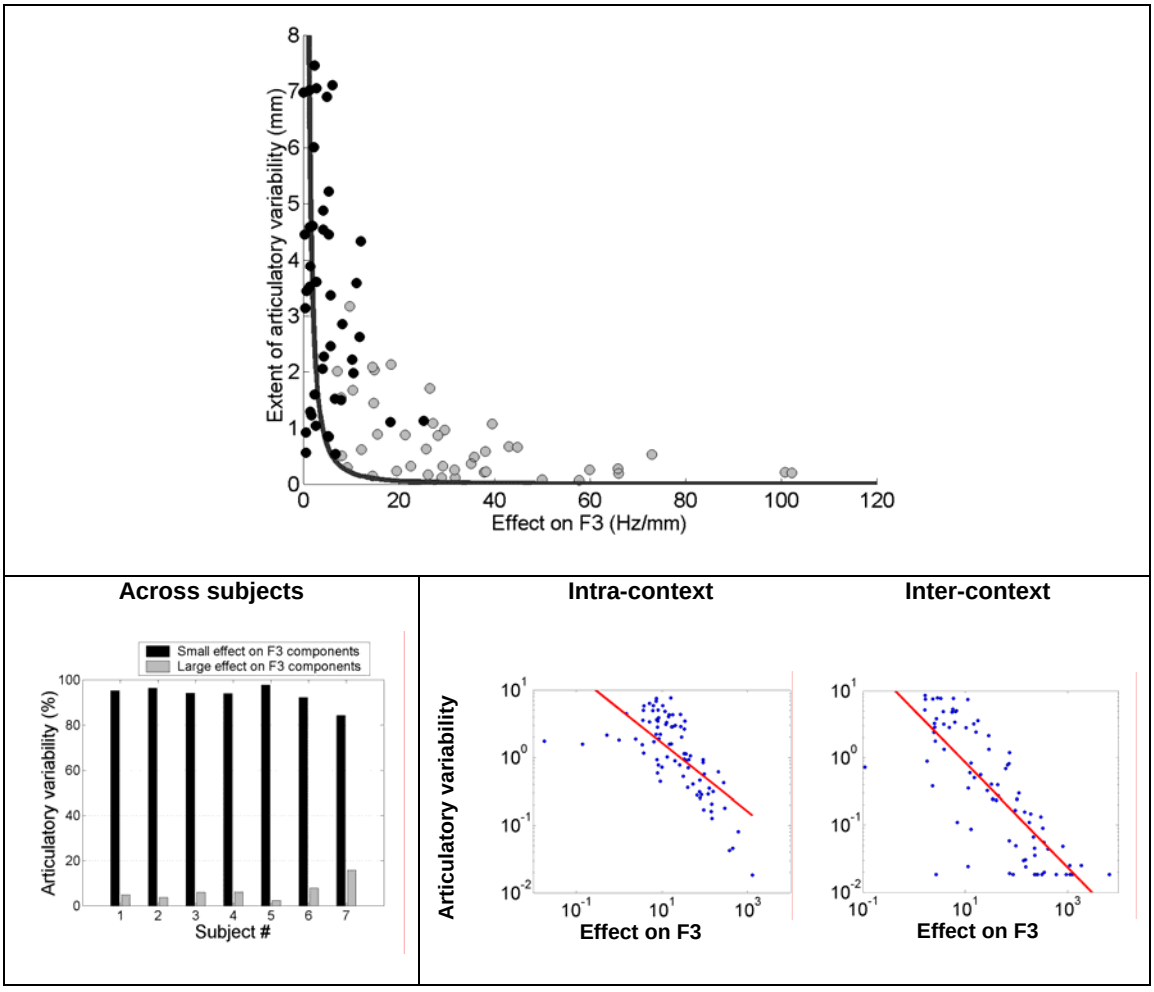

Figure 1.4. Experimentally measured articulatory/acoustic relations. **Top:** The extent of articulatory variability (in mm) vs. the effect on third formant frequency (F3, in Hertz) for all articulatory dimensions for all subjects (each dot represents an articulatory dimension - a direction of movement of the articulators - for a given subject). The thick line represents the inverse relation fit to data ($y = 4.10 \times ^{-0.58}$). The data fit was performed on the class of functions of the form $y = a * x^b$ fitted in the log plane (explaining why the fit looks biased toward the lower values in this linear plot). The 95% confidence intervals for the parameter values of this fit were a = [2.86, 5.90] ; b = [-0.72, -0.44]. The data fit shows a significant relationship between the tested variables ($F_{1,82} = 69.68$; $p=1*10^{-12}$). Black/gray points represent the articulatory dimensions that, for each subject, would be categorized as small/large effect on F3 components. The

inverse relation shown in this plot is identified as a *predictive relation between acoustic stability and articulatory variability* (i.e. the acoustic stability of a given articulatory dimension predicts the extent of articulatory variability each subject demonstrates along that articulatory dimension). **Bottom left:** Consistency of found articulatory/acoustic relations across subjects. The percentage of articulatory variance associated with large/small effect on F3 components is shown for each subject. Under the null hypothesis (articulatory variability not associated to the effect on F3) these percentages would be equal (50% each). Small/large-effect on F3 components are, from the 12 possible directions of articulatory variability measurable for each subject, the 6 directions that produce the least/most F3 change. A strong bias of the articulatory variability towards those articulatory dimensions that have a small effect on F3 is apparent in all the experimental subjects. **Bottom right:** Consistency of found articulatory/acoustic relations in terms of the intra- and inter- context articulatory variability. Context refers to the different leading phonetic contexts (/ar/ /dr/ /gr/ /vr/ /br/) in which the phoneme /r/ was produced. Intra-context articulatory variability is defined as that variability found in articulatory configurations when looking only at different productions in the same phonetic context. Inter-context variability is defined as the variability of the average articulatory configurations for each phonetic context. The plots show the observed relation between the effect on F3 and articulator variability acts similarly to reduce the acoustically relevant contextual articulatory variability (inter-context plot, R^2=.62), as well as the remaining acoustically relevant articulatory variability for each individual phonetic context (intra-context plot, R^2=.46). These data suggest that the observed articulatory/acoustic relations are the result of the action of a common motor control mechanism (one utilizing the same acoustic target for /r/ across phonetic contexts) rather than a context-dependent target definition.

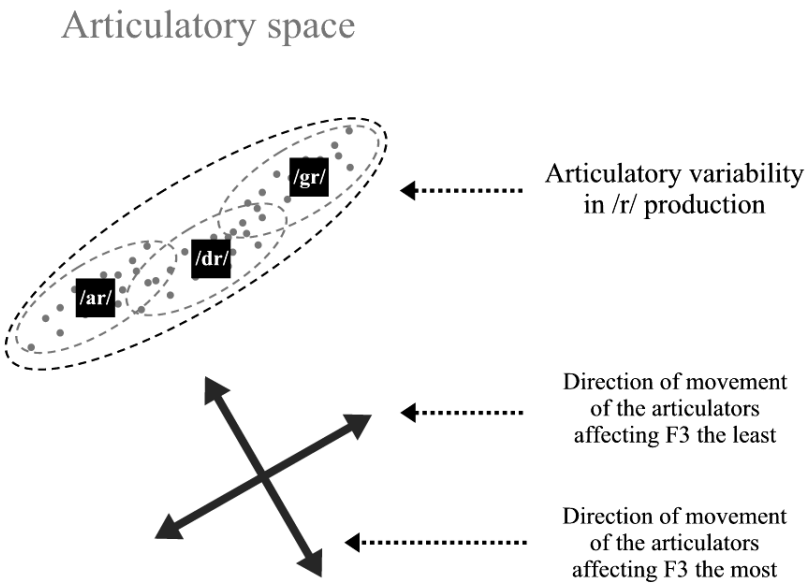

Figure 1.5. Diagram summarizing the main results in this section. The plot represents in a schematic way the range of articulatory configurations (dots in the plot) reached in the production of /r/ under different phonetic contexts (black boxes). The main results are: **a)** An acoustic variable (F3) is the best predictor among the phonemic target variables tested for the shape of the articulatory variability in the production of American English /r/. The articulatory variability is maximal along the directions of movement of the articulators associated with small F3 changes, and minimal along the directions of movement of the articulators associated with large F3 changes. **b)** The intra-context articulatory variability (the articulatory variability for each of the phonetic contexts) shows the same association with the effect of F3, indicating not the action of a context-dependent target definition, but possibly a common control mechanism utilizing an acoustic phonetic-target.

B. Speaker-specific vocal tract models

For the first two subjects participating in the previous analyses, we constructed from MRI and acoustic data a simple model characterizing the specificities of their vocal tracts and the range of

acoustic signals (limited to the first three formant values) that different configurations would produce.

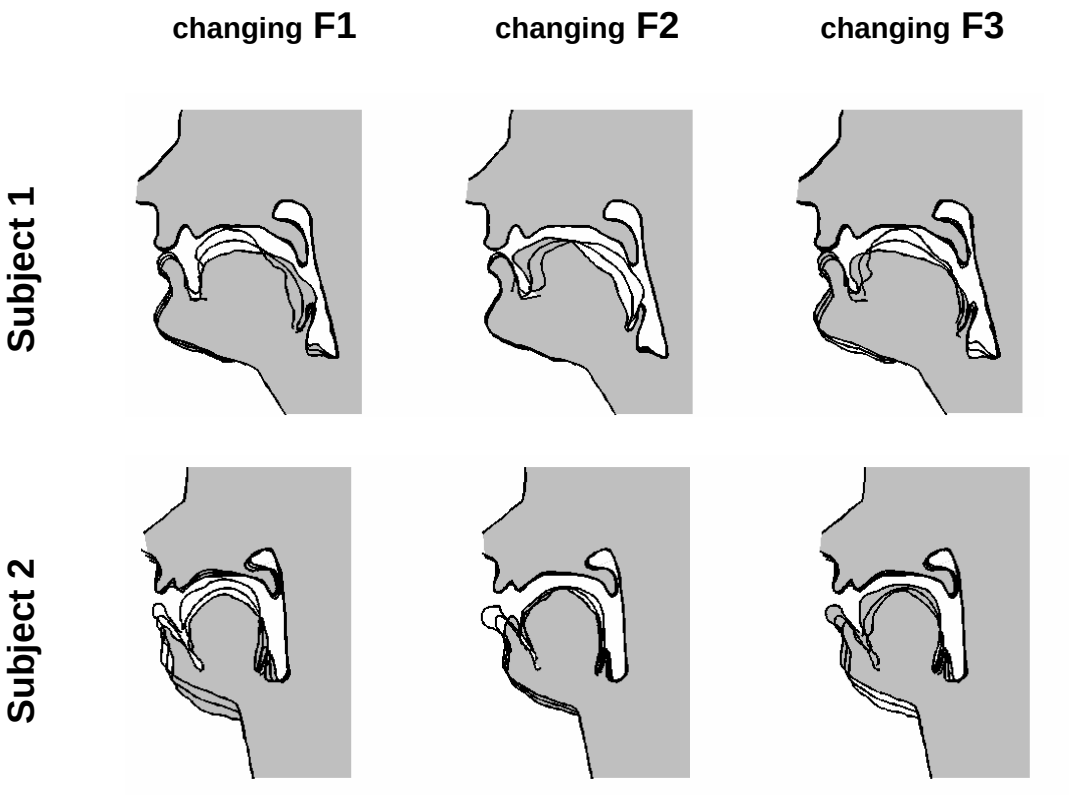

Figure 1.6. Sample movements of speaker-specific vocal tract models for Subjects 1 and 2 to change F1 (left), F2 (center), and F3 (right). For each subject, the deviations from a neutral articulatory configuration necessary to produce an individual change (increase/decrease) in each of the first three formants of the resulting auditory signal are shown (e.g., the first column represents the movements associated with changes in F1 while keeping F2 and F3 constant). The gray area represents the configuration that produces the highest formant value (for the corresponding formant) among the configurations represented.

A principal component analysis of the articulatory configurations led to a set of five meaningful articulatory components covering 75.4% and 83.7% of the total observed variability in shape for the two subjects, respectively. The jaw component primarily describes the aperture/closure of the mouth, along with the associated lip aperture/closure, and lowering/raising of the tongue body; the three tongue components describe approximately the raising/lowering of the apical and dorsal areas of the tongue and its front/back movement; the lip component describes the frontal extension (protrusion) of the lips (c.f. Maeda, 1990; see also the Discussion section). Components derived from other vocal tract segments (a velum component, describing the opening/closing of the nasal cavity; and a laryngeal component, describing the raising/lowering of the base of the laryngeal region), were estimated but not explicitly used in the simulations presented in this chapter (other than any of their movement that was associated with the jaw component). The articulatory to acoustic mapping was then estimated by a linear fit between the articulatory configurations (defined by the positions of each of these five components) and the corresponding acoustic output (defined by the first three formant values measured during the MRI scans). Figure 1.6 illustrates movements of the resulting speaker-specific vocal tract models to achieve changes in F1, F2, and F3. Each column represents for each subject the movement of the articulators, starting from a rest (average) configuration, that would be associated with changes in an individual formant. The results are consistent with standard characterizations (Schroeder, 1967; Fant, 1980) of high/low tongue configurations associated with low/high values of F1, respectively (left column in Figure 1.6), and front/back tongue configurations associated with high/low values of F2, respectively (middle column in Figure 1.6). At the same time, the resulting vocal tract models accommodate the specificities of each subject. For example, Subject 2 tended to use lip protrusion more actively to lower F2 (see for example Perkell et al, 1993, 1995, where trading relations between lip protrusion and tongue-body raising, argued to stem from their motor

equivalence in the control of F2, were investigated in the context of /u/ production). With respect to the action on F3, Subject 1's movement to increase F3 can be interpreted from an acoustic theory analysis as a decrease in the front cavity length together with an increase of the palatal constriction area, both acting to raise the third formant value, while Subject 2 appears to increase F3 primarily by decreasing the size of the front cavity.

C. Simulations of /r/ production

A simplified version of the DIVA model (Guenther et al., 1998) was used to control movements of the speaker-specific vocal tract models for Subjects 1 and 2 while performing /r/ productions in different phonetic contexts. An acoustic /r/ target was defined by its first three formants values ([593, 1238, 1709] Hz for Subject 1, and [376, 1476, 1990] Hz for Subject 2), and the simulations were run starting from articulatory configurations representative of the leading context phonemes (see Section II.C for details). In order to compare the model simulations to the EMMA data, approximate transducer locations were manually identified (see Methods section) on each subject-specific vocal tract model. Acoustic and articulator trajectories for the production of /r/ in the contexts /ar/, /dr/, and /gr/ were then obtained using the DIVA model. These contexts were chosen to represent the full range of articulations seen in the experimental data.

Figure 1.7 compares the experimentally measured EMMA data (first row) to the simulation results (second row) for each subject, in terms of the direction of movement of the tongue transducers. The initial transducer positions in the simulations is fixed to that obtained from the EMMA data 75 ms before the F3 minimum (dashed lines). The results indicate that the direction of movement estimated using the DIVA model for the three leading phonetic contexts closely approximates the experimentally measured data for both subjects. The correlation between modeled and experimental change in transducer positions (tongue gestures) was r=+0.86 and

r=+0.93 for Subjects 1 and 2, respectively. Qualitatively, the model mimics the range of /r/ configurations used by each subject in the phonetic contexts tested (thick black lines in Figure 1.7).

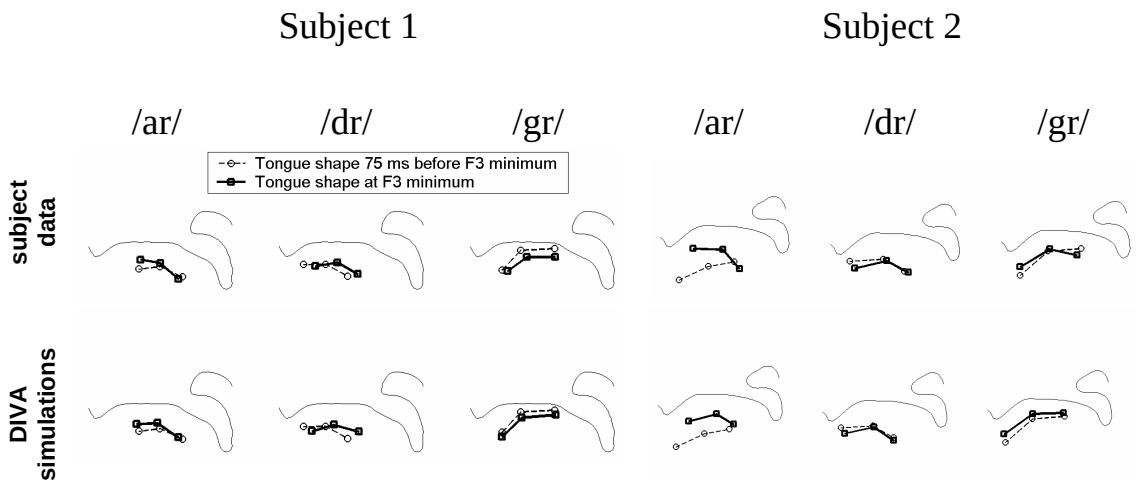

Figure 1.7. Simulations of the DIVA model producing /r/ in different leading phonetic contexts. Top row shows the average lingual gestures used by each subject when producing /r/ in the contexts (from left to right) /ar/, /dr/, and /gr/ as measured using electromagnetic midsagittal articulometry (EMMA). Bottom row shows the simulation results using the DIVA model (with a subject-specific acoustically defined target for /r/) in conjunction with each subject-specific vocal tract model. Dashed lines represent the initial (75 ms before F3 minimum) transducer positions, which are fixed to the experimentally observed values in the DIVA simulations. Solid lines represent the final transducer positions (at the F3 minimum for /r/). The outline of the hard palate and velum is included for reference. The correlation between the modeled and experimental movement of the tongue (tongue gestures) was r=+0.86 and r=+0.93 for Subjects 1 and 2, respectively.

Next we investigated the ability of an acoustic target speech motor control scheme to predict the emergence of the articulatory/acoustic relationship observed in the experimental data. To that end, we analyzed the /r/ production simulation final articulatory configurations when using a wide range of leading phonetic contexts. All available configurations from the MRI data of each subject were used as starting articulatory positions and the DIVA model was run using the same acoustic /r/ targets as in the preceding simulations. Analysis of the resulting articulatory variability led to the results shown in Figure 1.8. For each subject, the five articulatory dimensions show the expected predictive relations between acoustic stability and articulatory variability (Figure 1.8 left; cf. the experimental results in Figure 1.4 top). The relation between articulatory variability and effect on F3 predicted by the model is close to linear in the log variables (R^2=.91), justifying the use of this family of curves when fitting the experimental data (see Figure 1.3). For the simulated data, the linear regression on log variables shows a significant relationship between the tested variables despite the limited data ($F_{1,8}$ = 82.71 ; p=2*10^{-5}). As an additional test, we analyzed the initial articulatory variability (the variability of the contextual articulatory configurations, prior to any movement of the articulators) and confirmed that the articulatory/acoustic relation was not present in the contextual configurations prior to the action of the speech controller (p>0.39). This indicates that the relationship resulted from the movements produced by the DIVA model. Furthermore, the simulation results mimic the expected relationship as derived theoretically from the DIVA control equations (dotted line in Figure 1.8 left; see Appendix for this derivation). The nature of the inverse relation predicted by the model ($y \propto x^{-0.78}$) was slightly steeper than the one observed in the EMMA data ($y \propto x^{-0.58}$) but the confidence intervals for the two curve parameters overlap ([-0.98,-0.58] and [-0.72,-0.44], respectively). For completeness, Figure 1.8 (right) illustrates the consistency of articulatory/acoustic relations in the simulations across the two subjects (c.f. the experimental

results in Figure 1.4 bottom left). Overall, these results indicate that an acoustic target controller, such as the one used in the present simulations, predicts the relationship between acoustic stability and articulatory variability observed in the experimental data. Furthermore, the DIVA model produces articulatory movements that closely mimic those of a particular speaker when controlling a speaker-specific vocal tract model.

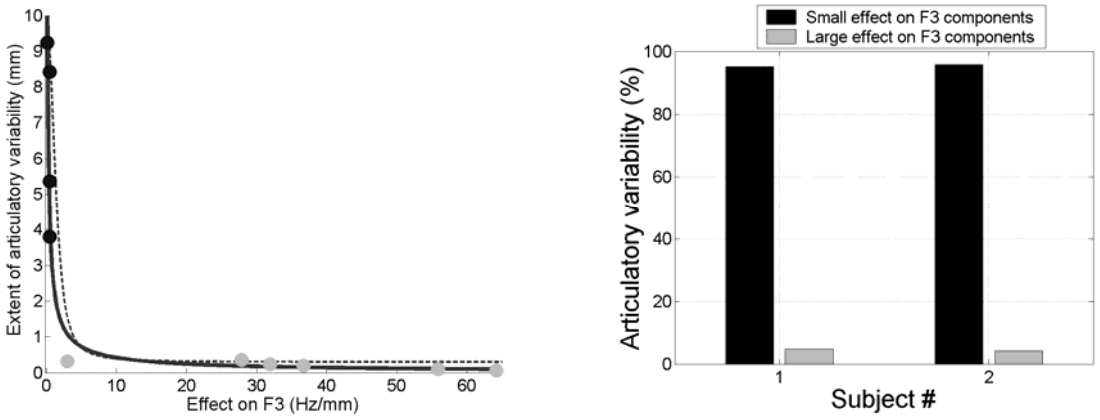

Figure 1.8. Simulated articulatory/acoustic relations in /r/ production using the DIVA model (compare to experimental relations in Figure 1.4). **Left:** The extent of final articulatory variability vs. the effect on F3 for all articulatory dimensions of both subjects' simulations. The solid curve represents the inverse relation fit to this data ($y = 2.54 \times ^{-0.78}$; equivalent linear fit on the log variables, R^2=.91). The 95% confidence intervals for the parameter values of the fit $y = a \times ^{b}$ were *a* = [1.47, 4.38] ; *b* = [-0.98, -0.58]. The estimated fit shows a significant relation between the tested variables ($F_{1,8} = 82.71$; $p=2*10^{-5}$). The dotted curve represents the expected predictive relation as theoretically derived from the DIVA model (see Appendix). Black/gray points represent the articulatory dimensions that would be categorized as small/large effect on F3 components, respectively. These results indicate that the control scheme proposed by the DIVA model predicts the emergence of articulatory/acoustic relations similar to those observed in the experimental data (see Figure 1.4 top). **Right:** Consistency of simulated articulatory/acoustic relations across subjects. The percentage of articulatory variance in the simulated /r/ productions associated with large/small effect on F3 components is shown for each subject. As in the experimental data (see Figure 1.4 bottom left), a strong bias of the articulatory variability toward those articulatory dimensions that have a small effect on F3 is apparent in both subjects' simulations.

IV. DISCUSSION

A. On coordinate frames and articulatory dimensions

In target-based speech motor control models, the question of what coordinate frame is used by each model is usually identified with the proposed target representation. The task-dynamic model of Saltzman and Munhall (1989) exemplifies a type of computational model that uses a vocal tract shape coordinate frame (vocal tract targets defined by tract variables). The DIVA model (Guenther et al, 1998) exemplifies a computational model that uses an acoustic coordinate frame (targets defined by acoustic variables). While there are many different coordinate frames one could use to represent the articulatory state, a major question for speech production modelers is what coordinate frame(s) provides a simpler or more parsimonious characterization of behavioral data. In the same way as physical laws can be more readily unveiled when using an appropriate coordinate frame (e.g. planet orbits from an earth-centered vs. a sun-centered coordinate frame), for speech production the use of an appropriate coordinate frame should allow the researcher to more clearly expose functional relations in the data. Finally, the ability of different coordinate frames to characterize the available motor speech production behavioral data could direct and facilitate the modeler's enterprise in proposing specific motor control strategies, and in particular it directly relates to the question of appropriate target definitions in target-based motor control schemes.

The behavioral data dealt with in this study is the articulatory variability present in American English /r/ production. Since articulatory variability is a local property (it characterizes the local departures in articulatory configurations from an average configuration) a linear approximation to the articulatory space geometry is appropriate. The issue of coordinate frames, under a linear approximation, becomes the simpler issue of characterization of vector spaces. Under this framework the articulatory space is a multi-dimensional vector space, and its characterization

reduces to the definition of an appropriate base (a set of independent articulatory dimensions, each describing a direction – or vector - in the articulatory space). Different bases would in this way characterize different coordinate frames for the description of the articulatory state. Each of the columns in Figure 1.6, for example, describes a different articulatory dimension (i.e. a direction of movement, or vector, in the articulatory space). The three articulatory dimensions in this figure characterize an acoustic coordinate frame (one based on three formant descriptors).

B. Predictive relations between acoustic stability and articulatory variability

A purely empirical approach to describing appropriate coordinate frames for the characterization of articulatory variability in /r/ production could be potentially given by a principal component analysis of the articulatory covariance. This analysis provides the set of independent articulatory dimensions that best (most simply) characterize the observed articulatory variability. Conceptually, these correspond to the articulatory dimensions that offer an optimal separability of the articulatory variability associated with each dimension. In a two-dimensional case, for example, the resulting two articulatory dimensions would correspond to those dimensions associated with the largest and smallest variability, respectively. A purely empirical approach like this, nevertheless, has potentially limited generalizability; i.e. since articulatory variability is a local property, the characterization resulting from the analysis of /r/ production might not be appropriate for other production examples. Furthermore, the researcher is left to interpret the resulting articulatory dimensions in terms of his/her theoretical constructs.

In this work we opted for a mixed empirical/theoretical characterization of the observed articulatory variability. In this way, we tested the ability of theoretically motivated articulatory dimensions to offer good separability of the observed variability in articulatory configurations. We feel that this approach has a better chance to generalize to other cases of speech production data, and that it offers a more useful source of information for the development of motor control

models of speech production. From this perspective, the relevance of the results presented in Figure 1.4 bottom-left is that they show how an articulatory dimension defined by an acoustic property (F3, a salient acoustic cue for /r/ perception), offers a good separability of the observed articulatory variability in /r/ production for all subjects tested. In particular, an average of 93% of the articulatory variability concentrates along articulatory dimensions that have a relatively small effect on the third formant (F3) value, while only 7% concentrates along articulatory dimensions which have a relatively large impact on F3. This result indicates that an acoustically-defined articulatory dimension would be a good candidate to enter an appropriate coordinate frame characterization of the presented speech production behavioral data. In the same way, these results suggest that motor control models utilizing an acoustic target representation can potentially provide a more parsimonious characterization of these behavioral data than models utilizing a different phonetic-target coordinate frame. Furthermore, following the original motivation for searching appropriate coordinate frame characterizations, we showed (Figure 1.3, and Figure 1.4 top) that using an acoustically defined coordinate frame can also be useful for unveiling functional relations in the behavioral data. In particular, we showed that the degree of articulatory variability associated with any particular articulatory dimension is related to the associated extent of change in F3 by a linear relationship in the log variables (R^2=.46; $p=1*10^{-12}$). This relationship is conceptualized as a predictive relation between acoustic stability and articulatory variability. The form of this relationship is again consistent with that expected from a control mechanism using an F3 target; i.e. the final articulatory variability is lower for those articulatory dimensions most relevant to determining the F3 value. We believe the ability of different speech motor control models to mimic the measured relationship between acoustic stability and articulatory variability could be a useful reference in creating and improving motor control models of speech production. In particular, this chapter argues that an acoustic target

model of speech production can account for the emergence of these articulatory/acoustic relations and the specificities of the measured articulatory variability.

An important issue regarding the observed articulatory/acoustic relations examines the extent to which they favor acoustic target motor control models in contrast to vocal tract target models. Several results of the present study build a very strong case for the acoustic target hypothesis. First, the results in Figure 1.3 indicate that while the tested acoustic variable (F3) shows a significant relation with the extent of articulatory variability, making it a potential candidate for a useful articulatory coordinate frame definition, the hypothesized vocal tract-variables fail to show such a relation. This negative result addresses what can also be observed from the temporal progression of tongue shapes. Figure 1.7-top shows the average tongue shapes adopted by Subjects 1 and 2 in different phonetic contexts. From the inspection of this data it seems, for example, that the movements of the tongue tip in different phonetic contexts for each subject do not seem to be aimed at any specific target configuration (see also Guenther et al., 1999). Observations such as this one are reflected in the negative results in Figure 1.3 regarding the tract variables tongue tip constriction degree and tongue tip constriction location. Another piece of comparative evidence between acoustic and vocal-tract target hypotheses addresses the possibility of context-dependent effects (context here refers to the phoneme preceding /r/). The results at the bottom right of Figure 1.4 indicate that the observed articulatory/acoustic relations do not solely stem from the context-dependent articulatory variability, and can be equally observed when focusing on the intra-context articulatory variability (i.e. the articulatory variability resulting from /r/ production in each specific phonetic context). This result again points towards hypotheses that posit the observed trading relations as resulting from the motor control strategy (such as the acoustic target hypothesis), rather than explanations that rely on context-dependent targets (such as the possibility of different articulatory targets for /r/). Last, the possibility of context-dependent

articulatory targets was also directly addressed by trying to show predictive relations between tract variables and intra-context articulatory variability. Our failure to observe such relations indicates that using context-dependent articulatory targets does not seem to significantly improve the predictive ability of hypothesized tract variables on the observed articulatory variability. Regarding the possibility of subject-dependent tract variable definitions (e.g. one subject might use a tongue tip constriction target, while another might use a tongue-dorsum constriction target) our analyses do not support this possibility. No pattern of consistently used tract variables emerges, and when an optimally defined tract variable is used for each subject, the average separability of the articulatory variability is still only 67% (compared to 93% of the acoustic hypothesis).

Overall, the results indicate that an acoustic frame of reference provides a more parsimonious characterization of the observed articulatory variability than an articulatory frame of reference (one defined by vocal tract constriction variables). One might argue that, given the linear nature of our analyses, articulatory targets defined as linear combinations of tract variables are completely equivalent to acoustic targets. From this perspective the results simply indicate that, if articulatory targets are being used, they are probably not defined by simple vocal tract constriction targets but could possibly be defined by non-trivial linear combinations of these variables. Even more specifically, they could be parsimoniously defined by those linear combinations that best relate to the effect on relevant acoustic cues, as exemplified by F3 in the current /r/ production data. Such targets would be in this case more simply characterized as acoustic.

C. Speaker-specific vocal tract models

The simulation results shown in this chapter also indicate that it is possible to construct simple speaker-specific vocal tract models approximating the specificities of each subject's speech production apparatus from a limited amount of MRI and acoustic data. To construct a parametric description of articulatory movements, most previous approaches (Perrier et al., 1992; Story et al., 1996, 1998) create a grid in the midsagittal plane and obtain the vocal tract area function from the intersection of this grid with the vocal tract cavity. Area functions are estimated from MRI data in order to approximate the vocal-tract articulatory/acoustic mapping (e.g. Maeda, 1990, using an elliptical approximation to the area functions; or Tiede and Yehia, 1996, using 3-D volumetric MRI representations of the vocal tract). Compared to these approaches, the proposed vocal tract model estimation requires a relatively small amount of MRI and acoustic data for each subject and does not require an appropriate estimation of the area functions (which poses technical difficulties, e.g. the teeth not being portrayed in MR images). The resulting speaker-specific vocal tract models presented in this chapter are in agreement with standard characterizations of articulatory to acoustic relations (such as the differences between high and low, front and back, tongue configurations predicted from perturbation theory, Schroeder, 1967; Fant, 1980) while accommodating the specificities of each subject's vocal tract and their effective articulatory degrees of freedom. The use of subject-specific vocal tract models, in conjunction with a speaker-independent motor control strategy, is a promising approach to fit the specificities of different subjects' speech movements.

The vocal tract models presented in this study use a restricted set of five articulatory degrees of freedom or dimensions: three for the tongue, and one each for the jaw and lips. The appropriate dimensionality of the articulatory space, or the functionally relevant subspace, is not yet ubiquitously agreed upon. Maeda's (1990) articulatory model proposes a seven-dimensional

articulatory space covering jaw, tongue, lips, and velum variability. In contrast, the Payan and Perrier (1997) biomechanical tongue model uses a set of seven muscle-related descriptors solely for the tongue articulatory space. The Rubin et al. (1981) articulatory synthesizer uses a ten-dimensional articulatory space covering jaw, tongue, lips, velum, and glottal variability, controlled by a nine-dimensional vocal tract space. The choice for the dimensionality of the speaker-specific vocal tract models presented in this chapter mainly reflects our interest in creating a simple speaker-specific vocal tract model using a limited amount of MRI data. Several factors limit the minimal amount of MRI and acoustic data necessary for the presented vocal tract models. Among the available MRI data, the subset corresponding to vowel and semivowel productions imposes the first limiting factor. These data are the ones being used in the linear articulatory to acoustic mapping, since they provide the clearest formant frequencies and are the ones where the linearity approximation is more justified. The number of vowel and semivowel production samples should at least exceed the number of articulatory degrees of freedom to provide a valid estimation. Given the high degree of linearity of the mapping between articulatory and acoustic descriptors present in this data, a number of vowel and semivowel samples just exceeding this minimum value seems to be sufficient (we used 9 and 6 configurations for Subjects 1 and 2, respectively). The total number of MRI articulatory configuration samples is the second limiting factor. These data are the ones being used for the principal component estimation of the articulatory degrees of freedom. A rule of thumb in principal component analysis is to use as many data samples as five times the number of simultaneous degrees of freedom to be extracted. In our study the tongue shape, with three degrees of freedom, provided the largest of the simultaneous degrees of freedom to be extracted (suggesting at least 15 vocal tract samples). For this estimation we used a set of 27 and 15 samples for Subjects 1 and 2 respectively. More complex models derived from larger datasets will arguably better describe each subject’s

effective articulatory degrees of freedom. Nevertheless, the presented models with five articulatory components account for a relatively large proportion of the articulatory configurations for each subject (75.4% and 83.7% of the variability in vocal tract configurations respectively). Furthermore, the estimated articulatory to acoustic mappings' agreement with standard conceptualizations of the articulatory to acoustic relationship reinforce our belief in the appropriateness of the resulting speaker-specific vocal tract models.

D. Acoustic target model predictions and simulations

Speech motor control models based on acoustic targets posit that the target for production of a phoneme is defined in terms of its acoustic properties, rather than as a specific vocal tract configuration. In this way the variability in articulator configurations in the production of a given phoneme would reflect the one-to-many relation between the acoustically defined target and the articulatory space (i.e. the range of articulator configurations that are able to produce sounds with equivalent acoustic properties). The DIVA model is an example of such a model. The simulations presented in this chapter use this model in conjunction with appropriate speaker-specific vocal tract models to replicate two of the subjects' articulatory data. The simulation results of /r/ production in different leading phonetic contexts (Figure 1.7 bottom) mimicked the range of articulatory gestures used by the two subjects being modeled (Figure 1.7 top). The correlation between the experimental and modeled tongue gestures was r=+0.86 and r=+0.93 for Subjects 1 and 2 respectively. Furthermore, the simulated articulatory configurations reached by the DIVA model showed similar articulatory/acoustic relations (Figure 1.8) as those found in the experimental data (Figure 1.4). In effect, the articulatory variability in the simulations along each articulatory dimension was inversely related to its associated effect on F3.

The ability of the DIVA model simulations to fit the specificities of each subject's lingual gestures for the characteristic phonetic contexts tested emphasizes the idea that a relatively wide

range of the articulatory variability in /r/ production can be explained by a simple speech motor control scheme using acoustic targets (without the need to appeal to possible multiple articulatory targets). In Figure 1.7-top, for example, the tongue tip for each of the subjects moves in different directions for each context, and these directions do not seem to aim at any common lingual configuration. Interestingly, this can be modeled simply as a movement in the articulatory direction that in each case brings the acoustic output closest to a fixed acoustic target. Similarly, as shown by the simulations, the same acoustic target model parsimoniously explains the emergence of predictive relations between acoustic stability and articulatory variability. The expected articulatory/acoustic relation theoretically derived from this model is exemplified in Figure 1.8 left (dotted line).

E. Limitations

There are several limitations of this study. First, the study is restricted to the analysis of American English /r/ production. The results presented could only be generalized if the motor control strategy used in speech production, which predicts the emergence of the observed articulatory/acoustic relations, is common across different phonemic targets. Evidence of articulatory trading relations argued to limit acoustic variability in the production of /u/ (Perkell et al. 1993) suggests another case where acoustic variables could potentially predict the extent of articulatory variability. It is thus likely that the descriptive ability of the acoustic-target hypothesis generalizes to other vowel and semivowel cases. Whether articulatory- or mixed articulatory/acoustic variables are more instrumental in the description of consonant productions is an issue that could potentially be addressed following a methodology similar to the one presented in this chapter. Our expectation would be that the exact nature of the phonemic targets (auditory and/or somatosensory) is learned, and it would depend on the amount of language- and subject- specific allowed variability in these two spaces for that phoneme. Second, the presented

articulatory/acoustic relation analyses are restricted to changes in F3. While this is an important acoustic cue for /r/ production, it is most probably not the only one. A more complex study showing the form of these relations when multiple acoustic cues are considered could potentially deepen our knowledge on the motor control strategies in speech production. In relation to this issue the simulations presented in this chapter use the first three formants as a descriptor of the acoustic /r/ target. The presence of a predictive relationship between F3 stability and articulatory variability in the simulations shows that for these relations to emerge it is not necessary for the targeted variable to be the sole descriptor of the target coordinate frame. Third, regarding the speaker-specific vocal tract models, the presented methodology is limited by the linear nature of the analyses involved. The relation between articulatory configurations and the acoustic output is complex. Nevertheless this relation seems to be well approximated by a linear relation between articulatory and formant descriptors if relatively open configurations (such as vowels and semivowels) are considered. In this way, the validation presented in the Methods section indicates that the appropriateness of the linear model extends for a relatively large proportion of the articulator space (as indicated by the good linear fits between articulatory and acoustic formant descriptors estimated using Maeda's realistic tube model). The proposed speaker-specific vocal tract models represent a simple first order approximation to the complexities of the vocal tract apparatus and the corresponding acoustic output. This approximation is especially valid for vowels and semivowels. For the production of consonants different strategies should be investigated. Finally, regarding the DIVA simulations, the small number of subjects modeled limits our faculty to generalize the model's ability to fit the specificities of each subject's articulatory configurations in different phonetic contexts. Our expectation would be that the inter-subject variability, assuming a speaker-independent motor control strategy, is mainly affected by differences in the subjects' vocal tract morphology, and hence could be accounted for by using

appropriate speaker-specific vocal tract models such as the one presented in this chapter. Future studies using speaker-specific vocal tract models could in this way help better understand the sources of inter-subject variability.

V. SUMMARY

The analysis of articulatory movement data on seven subjects during the production of American English /r/ in different phonetic contexts shows a functional relationship between acoustic stability and articulatory variability. This relation indicates that the extent of articulatory variability along any given articulatory dimension is well predicted by the effect that the articulatory dimension has on a relevant acoustic cue (F3): most of the articulatory variability present in the production of American English /r/ is concentrated along articulatory dimensions that produce minimal change in F3. Both the presence and direction of the observed relationship are consistent with speech motor control mechanisms utilizing an acoustic (F3) target representation. In contrast, no relationship was found between hypothesized vocal tract target representations and articulatory variability. The combined results indicate that if phonemic targets are being used, they do not seem to be simply defined by constriction variables, but as non-trivial linear combinations of them. Such variables are more parsimoniously defined in terms of an acoustic frame of reference.

The second part of this chapter investigated the ability of auditory or acoustic target models to explain the specificities of the range of articulatory gestures observed in the production of American English /r/. Speaker-specific models capturing the shapes of two subjects' vocal tracts were constructed from a combination of MRI and acoustic data. Simulations of the DIVA model (an example of an acoustic target motor control scheme) controlling each speaker-specific vocal tract model produced articulatory movements that closely mimic those of the speaker. Furthermore, the articulatory configurations realized by this model exhibit similar articulatory/acoustic relations as those observed in the experimental data. The results demonstrate the ability of motor control speech production models utilizing acoustic target representations to

mimic central aspects of the experimental articulatory data on a particular example of speech production.

CHAPTER 2. SPECTRAL ACOUSTIC TARGETS. REPRESENTATION OF ACOUSTIC EVENTS FOR SPEECH PRODUCTION.

I. INTRODUCTION

Chapter 1 introduced the notion of acoustic targets, and presented experimental evidence of the use of acoustic representations in speech motor control. In particular an acoustic representation based on formant descriptors in the context of the DIVA model was able to account for the experimental observations in a parsimonious manner. Nevertheless, a major pitfall of this work was that it was inherently limited to the case of vowels and semivowels, largely due to the inability of a formant representation to characterize the acoustic events occurring in many consonant productions. Furthermore, the extraction of formant descriptors is not robust to the presence of noise, and no parallel of a formant representation of the speech signal has been shown to occur in the brain areas involved in speech perception. All this leads to the current object of attention in this chapter, which is a modeling effort in the acoustic representation of speech that surpasses the limitations of the formant descriptors while inheriting some of its simplicity in the control of speech production.

Acoustic representations of the speech signal have been traditionally tackled from a speech perception perspective. This means that representations are judged on their ability to facilitate robust recognition of speech utterances. A main representative of this approach is cepstral coefficients (the Fourier coefficients of the log of the signal spectrum –the signal energy at different frequency bands), which have been shown to offer a useful representation in numerous

speech recognition applications. Those approaches starting from a speech production perspective have usually been limited to the analysis of the physical speech production apparatus. The main representative of this approach would be the LPC coefficients, which directly relate to an approximation of the vocal tract by a non-uniform tube model. The work presented in this chapter attempts to define an *acoustic representation for neural models of speech production*. This means that the focus will be on obtaining an acoustic representation that: 1) relates simply to the movement of the articulators; and 2) can be implemented by simple computations of modeled neurons in primary auditory cortex. Together with this we are interested in showing how the basic computations needed to control the speech production apparatus can be performed using acoustic targets based on the proposed acoustic representation.

We use the speech production framework of the DIVA model (Guenther 1998). Under this framework the speech articulators are controlled using a combination of feedforward and feedback commands. A feedforward motor command for each auditory target is learned from experience. During learning as well as during non-standard conditions (e.g. perturbation, bite-block, etc.) there will be a discrepancy between the desired acoustic target and the produced sound, and/or between the somatosensory expectations and the propiosensory information. A feedback command projects these error sources into an appropriate corrective motor command acting to reduce this discrepancy. This feedback loop implements an inverse control strategy for speech production. In this chapter we will focus on the auditory portion of this inverse control strategy. We will show how this control strategy can be implemented using spectral acoustic targets, and outline what are the necessary modifications to the present DIVA model in order to use spectral target definitions.

This chapter is organized as follows: Sections *I* and *II* sets the mathematical framework for the inverse control problem on spectral targets. Section *I* will analyze the articulatory-acoustic relationship and will propose a mathematical approximation relating the movement of the speech articulators to the associated changes in the sound. Section *II*, building on the previous model, will propose a novel measure of "acoustic difference" for comparing one sound to another; this measure is designed to be applied in the inverse control of the speech articulators. Section *III*, then, will take these mathematical models into the context of the representation of speech sounds in auditory cortex, and the DIVA model. Finally, examples of motor control using the proposed representations will be presented in Section *IV*.

A. The vocal tract

The vocal tract is the physical apparatus used to produce speech sounds. Figure 2.1 shows a simple schematic. Air originating from the lungs passes through the vocal folds and enters the vocal cavity. Depending on the state of the vocal fold muscles, the air may be forced to vibrate (for voiced sounds) or not (unvoiced sounds). The shape of the vocal cavity is largely affected by the configuration of the tongue, jaw, and lips. By changing its shape, speakers can produce a variety of sounds. Maeda (1990) derives a descriptor of the vocal tract shape based on seven parameters or dimensions: one parameter controlling the jaw aperture; two for the lips, controlling lip aperture and protrusion; three parameters affecting the tongue shape; and one controlling the larynx height. Three additional parameters define the source characteristics (the characteristics of the sound as it leaves the glottis). These parameters are defined as glottal pressure, glottal opening, and fundamental frequency, and together with the vocal tract shape descriptors can be used to estimate the characteristics of the produced sound for an arbitrary vocal tract configuration. We will denote this descriptor by a ten-dimensional vector **m**, which

characterizes both the vocal tract shape and source articulators. The elements of the vector **m** are defined to range between 0 and 1. The vocal tract then acts as a function **x**=f(**m**) transforming an arbitrary articulatory descriptor **m** into a sound, which we will denote by the yet unspecified vector **x**. The articulatory synthesizer of Maeda (1990) offers an approximation to this transformation for an average speaker. In contrast to other possible representations of the vocal tract state (for example the area function, which describes the area of the vocal tract cavity from the glottis to the lips) the descriptor **m** varies continuously and linearly with the movement of the vocal tract articulators. As discussed in Chapter 1 these are desirable properties for modeling control strategies of the speech articulators.

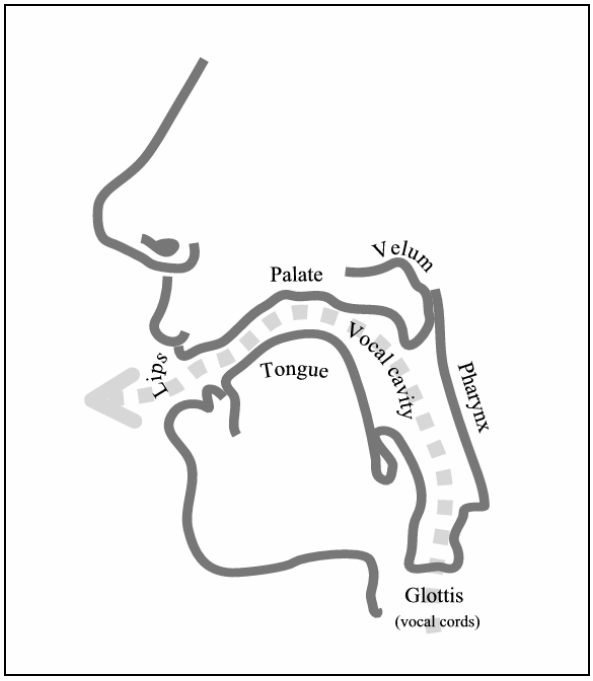

Figure 2.1 Vocal tract schematic. Air flows through the vocal cavity which is shaped by a number of articulators determining the produced sound characteristics.

B. Spectral shape

Sound consists of small fluctuations in air pressure. Often it is useful to imagine these fluctuations as a superposition of simpler waves oscillating at different frequencies. The sound spectrum represents the amount of air pressure energy at every possible frequency[2]. While human ears are sensitive to frequencies ranging from 20Hz to 20KHz, frequencies above 4 KHz are found to be less relevant to the intelligibility of vowels. Sounds are often highly non-stationary, and the spectrum of a complex sound such as speech is rapidly changing. Useful dynamic spectral measures for speech are usually obtained using short time-windows of approximately 40ms of length. In the human ear, oscillations in sound pressure are transmitted through the outer and middle ear to the cochlea. The cochlea consists of a tube about 3.5cm long curved into a spiral shape. Inside this tube, along the organ of Corti, lie thousands of hair cells. Each of these cells responds to the energy of the sound at a specific frequency (or limited ranges of nearby frequencies). Depending on their position along the cochlea, cells will respond to the low frequency (near the apex) or high frequency (near the base) components of the sound. This transition in sensitivity from low to high frequencies as we move along the cochlear tube is found to be roughly logarithmic. This is analogous to a roughly logarithmic scaling in human behavioral sensitivity to differences in perceived pitch of pure tones (mel scale; Stevens and Volkmann, 1940), and critical bands (Bark scale; Zwicker et al, 1957; Greenwood, 1961a, 1961b). We will denote by the vector **x** the log-spectrum of an arbitrary sound estimated at a discrete number of mel-spaced frequencies in a manner mimicking the cochlear cells' sensitivity. We will use this

[2] The spectrum is mathematically defined as the absolute value of the Fourier transform of the sound pressure wave form. For speech sounds, a closely related measure, the log-spectrum (i.e. the log of the spectrum), is preferred. The popularity of the log-spectrum over the simple spectrum results from the possibility of representing the spectrum as a linear mixture of different sources (corresponding to the glottal pulse, vocal tract transfer function, mouth radiation, transmission line distortion, etc.). In this chapter we will use the latter measure (log-spectrum) as a spectral characterization of speech sounds.

vector **x** as a sound descriptor, and also loosely as the modeled cochlear output[3]. We will use the notation **x**(t) when we want to emphasize the time-varying nature of the spectrum. Figure 2.2 exemplifies the modeled cochlear output **x**(t) following a sample speech utterance.

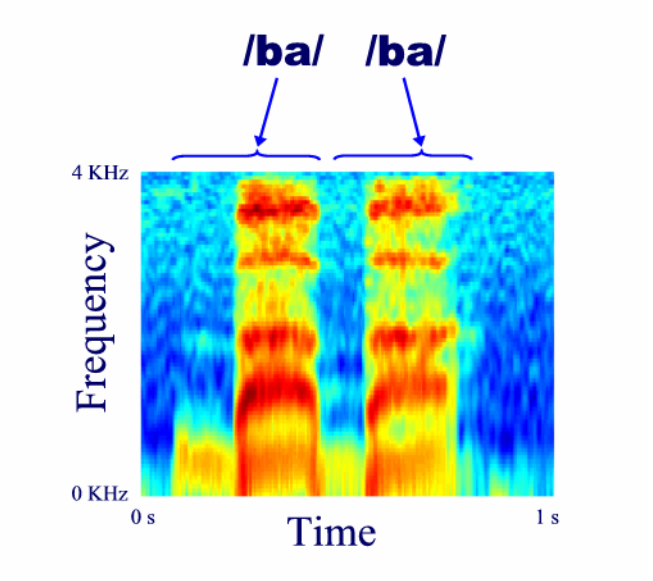

Figure 2.2 Modeled cochlear output **x**(*t*) following the production of the speech utterance /baba/. The cochlear output is modeled as the sound spectrum on a non-linear (mel space) frequency scale.

There are some aspects of the spectrum of speech sounds that can be characterized simply. One is its variation with sound amplitude (which is strongly related to the perceived intensity of the sound). The mean value of the spectrum (the average of the energy at all the frequencies) covaries with the scaling of the sound amplitude. The second aspect is its variation with respect to voicing characteristics (e.g. those distinguishing speaker identity) as well as common filtering resulting from the environment. These can be approximated as linear

[3] The details of our implementations are as follows: Sounds are sampled at 22050Hz and pre-emphasized using a 1st order filter with parameter a_1=-.95. The log absolute value of the short time-fourier transform of 2048 zero-padded samples is computed over windows of 40ms (using a hamming window), with 35ms overlap, and smoothed by low-pass filtering to the first 128 cepstral coefficients. The results are then re-sampled to 80 equally spaced points in a mel-scale between 0 and 4 KHz, and rectified (negative values are set to zero).

contributions to the sound spectrum, and are usually discarded or minimized in speech recognition algorithms using a simple linear projection. The remaining aspects of the spectrum of speech sounds are those actually most useful for the intelligibility of speech, and are the ones controlled by the vocal tract configuration. We will now turn to these aspects. In our formalism, we are interested in analyzing the variations of the sound spectrum **x** as we modify the articulatory configuration **m**.

C. Articulatory-acoustic mapping

There is a relatively simple, near linear relationship between the articulatory configurations of the vocal tract and the resulting formants of the produced sound, as described in Chapter 1. In this section we will attempt to parameterize the relation between the articulatory configurations **m** of the vocal tract and the spectrum **x** of the produced sound. Explicit vocal tract models (e.g. the Maeda synthesizer) provide an implementation of the articulatory-acoustic relationship but do not offer any insight into the nature of this relationship. The same can be said regarding artificial neural network approximations to this relationship (e.g. Blackburn 1996 used a multi-layer perceptron to approximate the relationship between vocal tract shape and log spectral sound descriptors in the context of a speech recognition system). In contrast with these approaches, this section will offer a mathematical parameterization of the articulatory-acoustic relationship. This mathematical description will be used in Section II to derive appropriate control strategies for the inverse control problem in speech. In particular, the results in this section will demonstrate two points: a) that the articulatory-acoustic relationship between articulatory configurations **m** of the vocal tract and the log-spectrum **x** is highly non-linear; and b) that this relationship can be efficiently approximated by a specific form of non-linearity in the context of multi-parametric groups.

We will start with an example of the cochlear output resulting from a linear movement of the articulators from one articulatory configuration $\mathbf{m_1}$ to a second configuration $\mathbf{m_2}$. In this case $\mathbf{m}$ takes the simple form:

$$\mathbf{m}(\lambda) = (1-\lambda)\cdot \mathbf{m}_1 + \lambda \cdot \mathbf{m}_2$$

Eq. 2. 1.

where λ is an arbitrary parameter varying between zero and one representing the different stages in the transition from the first configuration $\mathbf{m_1}$ to the second $\mathbf{m_2}$[4]. Using the articulatory synthesizer as an approximation of the transformation $\mathbf{x}=f(\mathbf{m})$ we can estimate the sound spectrum $\mathbf{x}(\lambda)$ associated with this articulatory movement. Figure 2.3 shows the spectrum of the produced sound.

[4] While the parameter lambda can be interpreted in the context of the *lambda model* (Feldman 1966) the present work does not address or assume any specific strategy for the muscular level control of the articulators. Rather this parameter is herein simply used to represent an arbitrary motor control variable parameterizing a range of articulatory configurations.

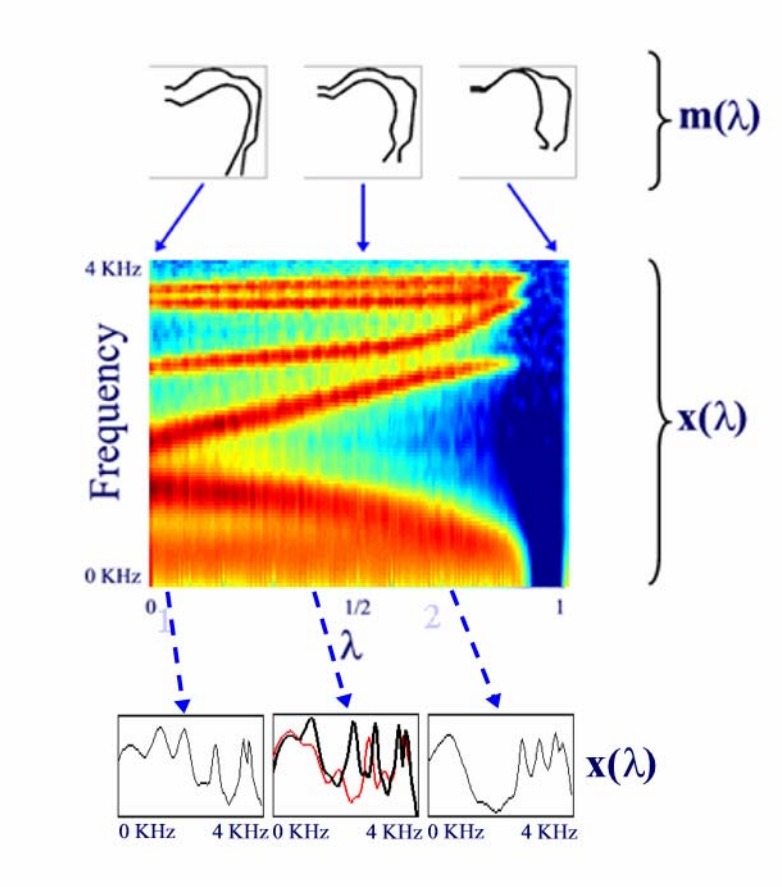

Figure 2.3 Cochlear output **x**(λ) following a speech movement characterized by an articulatory trajectory **m**(λ) of the form shown in Eq. 2.1. **Top**: Three articulatory configurations along the articulatory trajectory **m**(λ), corresponding to the points λ=0, λ=1/2, and λ=1, respectively. Each plot represents the vocal cavity outline (lips facing left) at each articulatory configuration. **Middle**: Acoustic output **x**(λ) produced by the modeled vocal tract as the articulators move along the trajectory **m**(λ). **Bottom**: Acoustic output **x**(λ) at three representative points during the non-silent portion of the production. For the intermediate configuration the vocal tract acoustic output (thick line) is compared with the average of the acoustic output at the two more extreme configurations (thin line). This plot highlights the non-linear nature of the acoustic trajectory plotted above.

We will compare two alternative mathematical models approximating the signal **x**(λ). The first (Eq. 2.2a and Eq. 2.3a) is a linear approximation, and it represents the current form in the DIVA model in which the articulatory-acoustic mapping is locally approximated (as implemented using a hyperplane radial basis function –HRBF- network; see Guenther et al., 1998 for details). The second is the one proposed in this dissertation, and represents an exponential approximation of the form shown in Eq. 2.2b and Eq. 2.3b.

Linear model

$$\mathbf{x}(\lambda) = \mathbf{x}_1 + \lambda \cdot \mathbf{h} \qquad \textbf{Eq. 2.2a}$$

$$\frac{d}{d\lambda}\mathbf{x} = \mathbf{h} \qquad \textbf{Eq. 2.3a}$$

Exponential model

$$\mathbf{x}(\lambda) = e^{\mathbf{H}\cdot\lambda} \cdot \mathbf{x}_1 \qquad \textbf{Eq. 2.2b}$$

$$\frac{d}{d\lambda}\mathbf{x} = \mathbf{H} \cdot \mathbf{x} \qquad \textbf{Eq. 2.3b}$$

The parameters **h** and **H** characterize in each model the direction of change in the spectrum **x** as the control parameter λ is modified, and they are expected to depend on both $\mathbf{m_1}$ and $\mathbf{m_2}$. These parameters take the form of a vector (**h**) and a matrix (**H**), respectively. In both cases these parameters can be estimated from Equation 2.3 using linear-regression techniques (see Appendix II.A for the description of a regression technique that allows the estimation of the parameter **H** in the exponential model from sample data $\frac{d}{d\lambda}\mathbf{x}$ and **x**). Both models provide a local approximation or fit to the changes in the sound spectrum **x** when varying the control parameter λ (Eq. 2.3). The total number of free parameters, or complexity, is larger for the exponential model than for the linear model[5]. In the *linear model* the parameter **h** is a vector with 80 elements

[5] For a short introduction to matrix analysis concepts useful to understanding the mathematics of the exponential model see Appendix II.A.

(one for each frequency band). Its elements can be interpreted as rate of change of energy at each frequency band. For example, a positive value of 0.1 in the element h_i means that the energy at the i-th frequency band is increasing during the modeled acoustic trajectory (at a rate of 0.1 dB s^{-1}). In the *exponential model* the parameter **H** is a matrix with 80x80 elements (one for each frequency band pair). Its elements can be interpreted as modeling the energy transfer between pairs of energy bands. For example, a positive value of 0.1 in the element H_{ij} means that the i-th frequency band is increasing its energy at a rate proportional to the energy at the j-th frequency band (at a rate of 10% s^{-1}). Positive/negative values in the diagonal element H_{ii} will be associated with exponential increases/decreases of the energy at the i-th frequency band. Positive and negative values along the off-diagonal element H_{ij} will be associated with energy transfers between the i-th and j-th frequency bands.

On a descriptive level, the changes in the spectrum seem to be well described as local frequency shifts. Those readers familiar with the mathematical description of frequency shifts will recognize that these are highly non-linear transformations, and probably will foresee severe limitations of the linear model to appropriately describe these changes. Also it is important to note that the space of sound spectra produced by the vocal tract is non-convex. This means that a linear interpolation between two valid spectra (those resulting from sound produced by the vocal tract) often results in a non-valid spectrum (one that cannot be produced by the vocal tract). This is suggested by the example shown in Figure 2.3 bottom. In these plots a linear interpolation between two valid spectra leads to an average spectra that not only considerably differs from a realistic intermediate spectrum, but it is also likely not achievable as the acoustic output at any articulatory configuration. Again, this non-convexity could pose serious limitations to the direct application of linear mapping techniques. The exponential model does not assume convexity or

linearly interpolable trajectories, but instead uses its increased complexity to approximate realistic interpolations from the observed acoustic trajectories. These observations predict a relatively poorer performance of the linear model. To validate them quantitatively and to learn to what extent the exponential model could outperform a linear model, we estimated for each model the extent of their local approximations (i.e. how local these local approximations are). We compared the two model fits (Equations 2.3a,b) based on their robustness to simultaneous equal perturbations of the original articulatory configurations $\mathbf{m_1}$ and $\mathbf{m_2}$. Note that if we equally perturb $\mathbf{m_1}$ and $\mathbf{m_2}$ the resulting articulatory trajectory will be nearby and parallel to the original $\mathbf{m}(\lambda)$. The direction of change of the articulators is unchanged by these perturbations, and at least for the extent of validity of these local approximations one would expect the direction of change of the spectrum to also remain roughly constant. The extent of articulatory perturbations for which the local approximations offered by each model remain at equivalent fit-levels was estimated. Figure 2.4 shows the sizes of these regions (in percentage of the total articulatory space), when varying the R^2 fit-level threshold. The linear model drops below a $R^2>0.9$ fit-level for articulatory perturbations larger than 0.00003% of the total articulatory space. The exponential model, at the same R^2 fit-level, permits perturbations of as much as 13% of the articulatory space. As this plot exemplifies, the extent of validity of the local approximation provided by the exponential model is considerably larger than the one provided by the linear model.

In order to emphasize the relevance of these results, they can be interpreted in terms of the number of local fits, or centroids in an RBF network, that would be necessary to implement an articulatory-acoustic mapping able to predict the direction of change of the spectrum $\mathbf{x}$ as we move the articulators in the fixed direction $\mathbf{m_2}$-$\mathbf{m_1}$. The number of centroids would be 3,286,965

and 8 for the linear and exponential model, respectively. This means that the additional complexity of the exponential model with respect to the linear model is well compensated by its associated increase in validity.[6] Overall, the analyses shown in this section indicate that: *1)* when considering a spectral representation of the acoustic signal, the articulatory-acoustic mapping is highly non-linear; and *2)* this non-linearity can be well approximated by the proposed exponential model.

The exponential model introduced in this section offers a modeling approximation to the relationship between articulatory parameters and the sound spectrum (or cochlear output). It acknowledges the strong non-linearities present in this relationship, and provides a simple way to model them. In the next section we will propose a strategy for the inverse control of the speech articulators. This strategy will be based on a simple gradient descent technique modified to better account for the characteristic non-linearity of the exponential model.

[6] We repeated all these computations using different starting articulatory configurations $\mathbf{m_1}$ and $\mathbf{m_2}$, to make sure the results were not specific to the original random choice of these parameters.

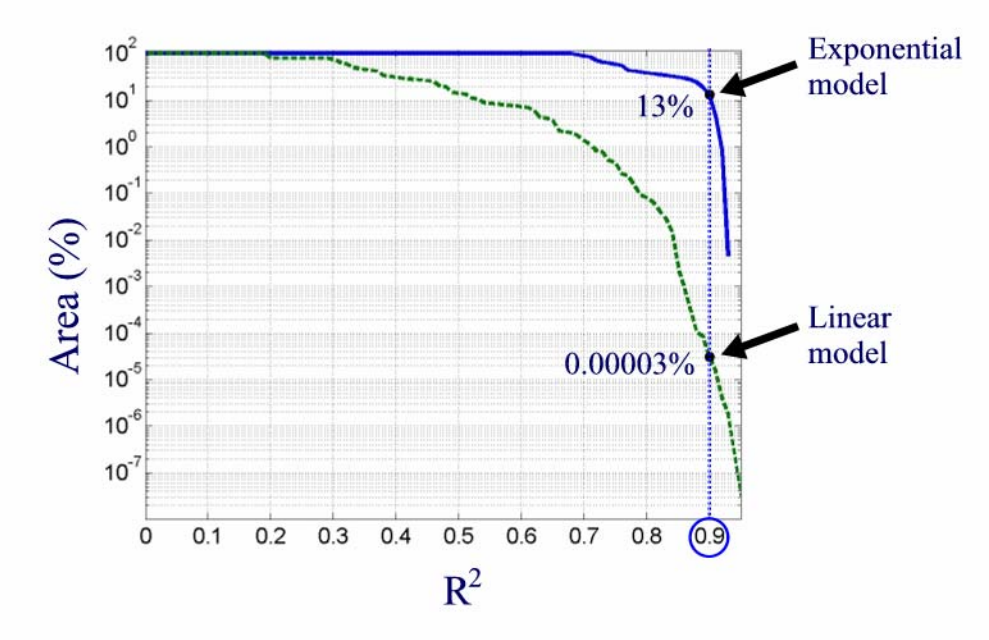

Figure 2.4 Validity regions for the linear (dashed line) and exponential (solid line) models, locally approximating the articulatory-acoustic data **x**(λ) centered at the example shown in Figure 2.2. This plot represents the area (in terms of the percentage of the total articulatory space) covered by each local approximation at equivalent fit levels (R^2). For example, the exponential model provides an R^2>.9 fit to the modeled articulatory-acoustic data for a region extending to 13% of the articulatory space, while a linear model at the same R^2>.9 fit-level would only extend to 0.00003% of the articulatory space.

II. CONTROL STRATEGIES

A. Uni-parametric control

We will start by assuming that we wish to control a single articulatory parameter λ whose effect on the spectrum **x** can be parameterized by Equation 2.3b, which we replicate here for convenience:

$$\frac{d}{d\lambda}\mathbf{x} = \mathbf{H} \cdot \mathbf{x}$$

Eq. 2.3b

Note that λ parameterized in the previous section an articulatory trajectory, and it was fixed in order to estimate the matrices **H** characterizing the acoustic consequences of this movement. In contrast, in this section we will do the opposite: we will use λ as a freely controllable articulatory parameter, and consider **H** given (its associated effect on the acoustics), in order to obtain appropriate control strategies on spectral targets. We will assume that the matrix **H** in Equation 2.3b is anti-symmetric. This matrix is also real-valued, as the vectors **x** are also always real. The anti-symmetry assumption has the effect of stabilizing the associated transformation $\mathbf{e}^{\mathbf{H}}$ by making it norm-conserving (i.e. the norm of the vector $\mathbf{e}^{\mathbf{H}\lambda}\,\mathbf{x}$ is the same as the norm of **x**, for any arbitrary λ). This normalization also represents our interest in the "shape" of **x** (the sound spectrum) independent of its norm (which is mainly influenced by the sound amplitude). For the analyses in this section we will assume that the vectors **x** representing the sound spectrum are normalized (of unit norm). In Section III we will re-introduce the sound amplitude into the control strategies.

Motivated by the discussion in Chapter I we will use a simple inverse-control strategy for controlling λ. This means that given a target spectrum $\mathbf{x}_{\text{target}}$ we would like to control the parameter λ using an equation of the form:

$$\frac{d}{dt}\lambda(t) = G\left(\mathbf{x}_{\text{target}}, \mathbf{x}(\lambda)\right)$$

Eq. 2.4

where *G* represents a function of the target ($\mathbf{x}_{\text{target}}$) and present ($\mathbf{x}(\lambda)$) spectra, and can be identified with a corrective motor command. We wish to find a suitable function *G* that will make $\mathbf{x}(\lambda)$ tend towards $\mathbf{x}_{\text{target}}$ (this is called a proportional control strategy). We will interpret this inverse control strategy in a discrete manner. Under the exponential model the set of possible acoustic states that we can reach starting from $\mathbf{x}(\lambda)$ and applying a discrete change *G* to the control parameter λ, forms a uni-parametric group of the form:

$$\mathbf{x}\left(\lambda + G\right) = e^{\mathbf{H}\cdot G} \cdot \mathbf{x}(\lambda)$$

Ideally we would like to affect the articulator λ by a quantity *G* that would make $\mathbf{x}(\lambda+G)$ most similar to $\mathbf{x}_{\text{target}}$. That is, we want to perform an "inversion" of the previous equation. This in the linear case would be performed by a simple matrix inversion (or pseudo-inverse). In the case of the exponential model, we found the following function to provide a good approximation to this "inversion" problem (see Appendix II.B for details on this derivation):

$$G\left(\mathbf{x}_{\text{target}}, \mathbf{x}(\lambda)\right) = \mathbf{x}_{\text{target}}^{\text{t}} \cdot \delta(\mathbf{H}) \cdot \mathbf{x}(\lambda)$$

Eq. 2.5

where the function δ(z) is defined from the derivative of the *sinc* function as:

$$\delta(z) \equiv \frac{d}{dz}\text{sinc}\left(\frac{z}{j\pi}\right)$$

Eq. 2.6

Note that this definition is slightly different from what we would reach using a gradient-descent approach to the inverse-control problem using the exponential model approximation. In this case we would obtain a function *G* of the form:

$$G\left(\mathbf{x}_{\text{target}}, \mathbf{x}(\lambda)\right) = \mathbf{x}_{\text{target}}^{\text{t}} \cdot \mathbf{H} \cdot \mathbf{x}(\lambda)$$

Eq. 2.7

Given the non-linear nature of the exponential model, a gradient descent technique is more likely to be affected by local minima. We would then expect that the proposed function *G* based on the approximate inversion of the exponential model will behave similarly to the gradient-descent result for small differences between $\mathbf{x}(\lambda)$ and $\mathbf{x}_{\text{target}}$, but will offer better results for large differences between the present and target spectra. This is confirmed by the results shown in Figure 2.5. This figure shows the correlation between the function $G(\mathbf{x}_{\text{target}}, \mathbf{x}_1)$ and the unknown parameter *k*, the ideal corrective command that would bring the present state towards the target. The spectra $\mathbf{x}_1$ and $\mathbf{x}_{\text{target}}$ are related through the equation $\mathbf{x}_{\mathbf{target}}=e^{\mathbf{H}\,k}\,\mathbf{x}_1$.

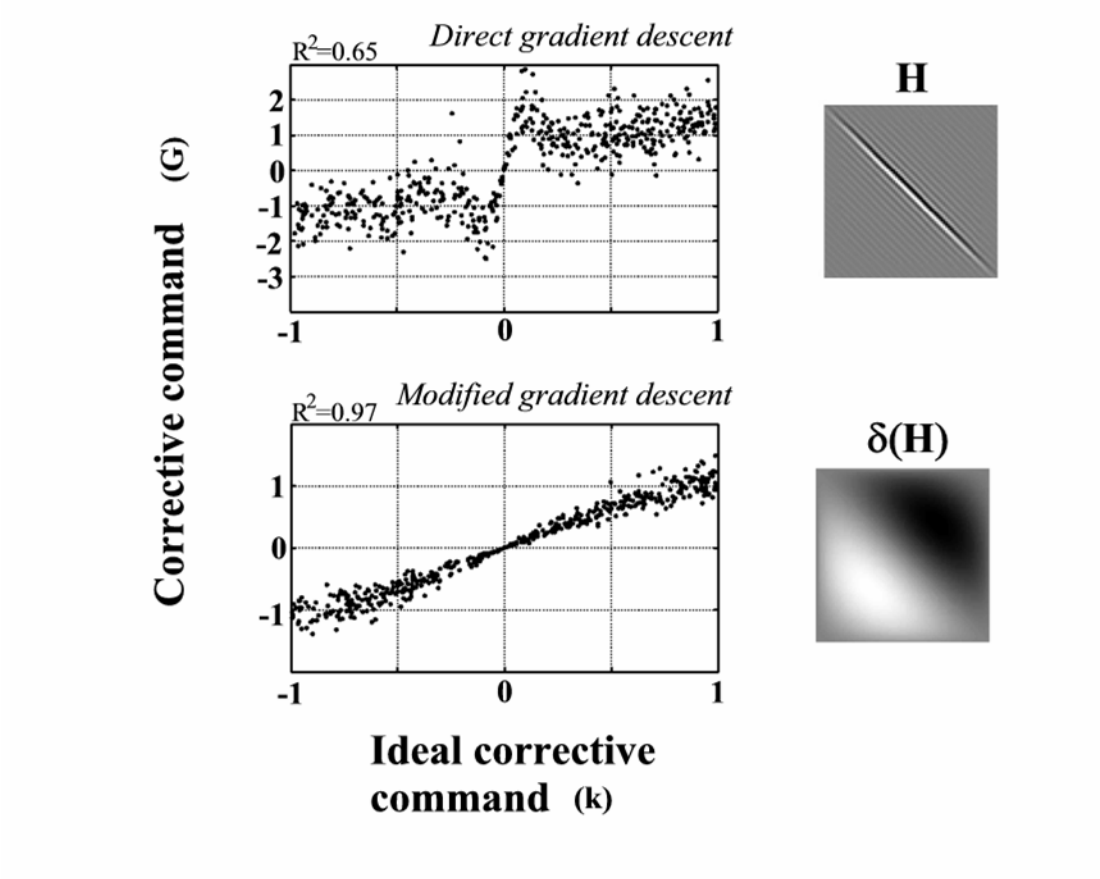

Figure 2.5 Comparison of the proposed control strategy (bottom) to a direct gradient descent technique (top). If the target $\mathbf{x}_{target}$ and the present state $\mathbf{x}$ are related by an exponential model of the form $\mathbf{x}_{target} = \mathbf{e}^{\mathbf{H}k}\mathbf{x}$, an inverse control technique computes a corrective command G from $\mathbf{x}$ and $\mathbf{x}_{target}$ that would bring the present state towards the target. Ideally the command G should approximate the value k. We defined randomly the values of the parameter k and estimated corrective commands using standard and modified gradient descent techniques (see text for details). The left plots show the relation between the parameter k and the estimated corrective commands (G). The modified gradient descent technique better approximates the ideal corrective commands as a function of the present and target states. In these simulations $\mathbf{H}$ is defined as that producing a spectral shift over all the frequencies of the spectrum (saturating at the 0 and 4KHz). The matrix $\mathbf{H}$ is shown on the right top plot. The matrix $\delta(\mathbf{H})$ used in the modified gradient descent algorithm is shown in the right bottom plot.

The proposed function *G* (shown on the bottom of Figure 2.5, as “modified gradient technique”) offers an improved estimation of the unknown parameter *k* (R^2=.97 vs. R^2=.65), especially for large values of this parameter (corresponding to large differences between $\mathbf{x}_1$ and $\mathbf{x}_{target}$). In this example we have used a matrix **H** defining a global spectral shift over all the frequencies of the spectrum (saturating at 0 and 4 KHz). This matrix takes the form shown in Figure 2.5 top right. The vector $\mathbf{x_1}$ was defined by smoothing a random Gaussian vector, and the parameter *k* was defined to range from -1 to 1.

Figure 2.6 shows four examples of the proposed inverse control strategy acting on a single articulatory parameter λ. For simplicity we assume this parameter to control a hypothetical articulatory dimension that produces the same global spectral shift in the sound spectrum as defined above. The plots in the top show two examples of inverse control. The spectral targets for these examples are defined as a random shift from the starting spectra. As these plots indicate, the proposed inverse control technique is able to smoothly reach the target spectrum (dashed lines in plots labeled as final state) from a relatively distant initial spectral configuration (plots labeled initial state). The plots in the bottom show two examples where the spectral target is defined randomly and lies outside the range of possible productions in our example (i.e. those sounds that can not be defined as simple spectral shifts from the original sound spectrum). These plots indicate that the proposed inverse control strategy performs robustly in these cases, converging, through pure spectral shifts, towards a state where the produced sound spectrum roughly approximates the desired target.

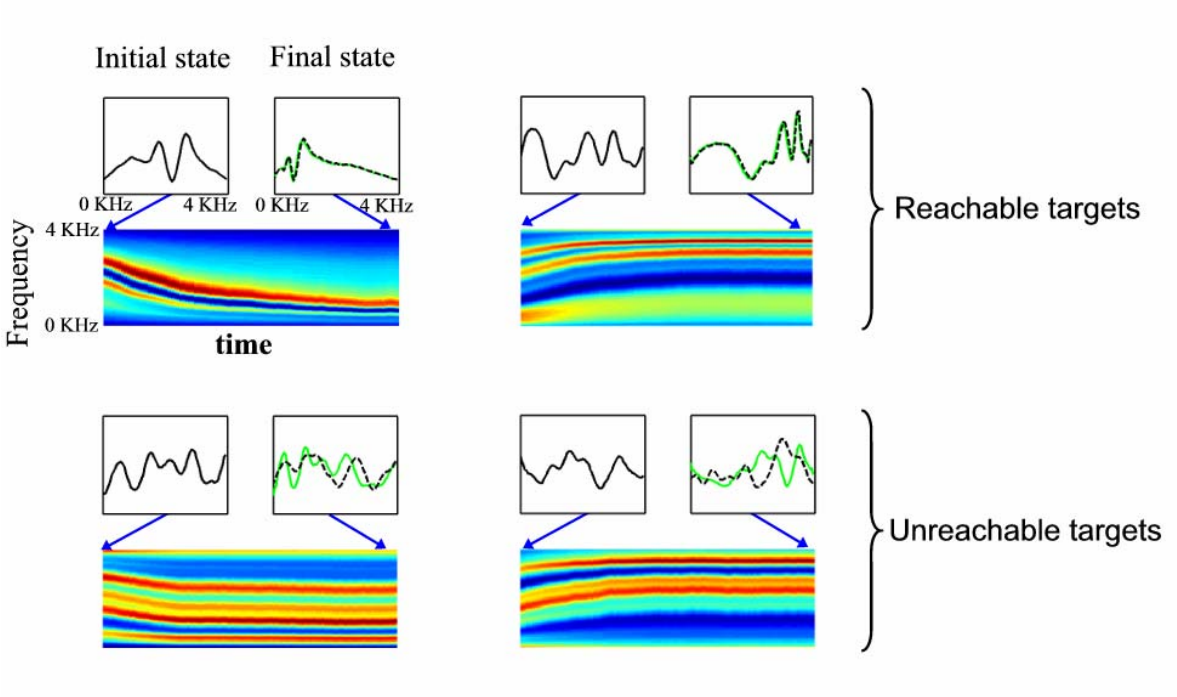

Figure 2.6 Four examples of inverse control on a single hypothetical articulatory parameter controlling a global frequency shift over the entire range of the measured acoustic spectrum (0 to 4KHz). **Top**: Inverse control on spectral targets that are possible to reach using the controllable articulatory parameter. Initial state plots show the sound spectrum at the initial configuration where the inverse control is started. Final state plots overlay the final state (solid line) with the target spectrum (dashed line) after 100 discrete iterations of the inverse control strategy. Images show the dynamic spectrum over the 100 iterations (axis labeled time) indicating a smooth convergence from the initial spectrum towards the target spectrum. When the spectral target lies within the range of possible productions, the proposed control strategy reaches a final state where the produced sound spectrum equals the desired spectral target. **Bottom:** Inverse control on spectral targets that cannot be reached using the controllable articulatory parameter. When the defined spectral targets are outside the range of possible productions, the inverse control strategy still converges towards a state where the produced sound spectrum (solid lines in final state plots) roughly approximates the spectral target (dashed lines in final state plots).

B. Multi-parametric control and the exponential difference measure

In this section we will extend the notions introduced in the context of uni-parametric control (controlling a single articulatory parameter) to the case of multi-parametric control (controlling a number of possibly redundant articulators). The exponential model offers a uni-parametric approximation to the spectrum when moving the articulators in a given direction $\mathbf{m_2}$-$\mathbf{m_1}$. Its natural multi-parametric extension to arbitrary articulatory directions $\mathbf{\Lambda}$ would be:

$$\mathbf{x}(\mathbf{\Lambda}) = e^{\sum_i \mathbf{H}_i \cdot \Lambda_i} \cdot \mathbf{x}_1$$

Eq. 2.8

where the vector $\mathbf{\Lambda}$ now represents the direction and extent of movement of the articulatory configuration **m**, and **x**($\mathbf{\Lambda}$) (the set of possible spectral states that could be reached from the current spectral state $\mathbf{x_1}$) now defines a multi-parameter group. In this equation each element Λ_i (the elements of the vector $\mathbf{\Lambda}$) represents the extent of movement along a given articulatory direction, and the corresponding matrices $\mathbf{H_i}$ characterize the spectral changes associated with each of these individual movements. For any arbitrary direction $\mathbf{\Lambda}$ this equation reduces to Eq. 2.2b, where the corresponding matrix **H** would be defined as $\sum_i \mathbf{H}_i \cdot \Lambda_i$.

A rigorous group-theoretically mathematical approach to this extended definition of the exponential model is beyond the scope of the current project. Selig (1996) offers an introduction to the use of multi-parametric groups in the context of robotics. It is interesting to note that the action of a multi-joint robotic arm can in fact be effectively characterized using equation 2.8 (for this case, the vector **x** would represent the spatial position of the arm end-effector, the elements Λ_i

represent the multiple joint angles, and the matrices $\mathbf{H_i}$ characterize the action of each joint). While these methods have proven successful in the characterization and solution of complex problems in robotics, in particular regarding robotic design and construction, the issue of inverse kinematics remains particularly challenging. Even for relatively reduced problems involving a few articulators (e.g. 6 joints of a robotic arm) and a few target dimensions (e.g. 3 spatial dimensions on a reaching task) it has not been possible to date to use these methods to derive a general solution to the inverse kinematics problem (Selig 1996). Furthermore note that, compared to this, the dimensionality of the speech control problem is an order of magnitude larger (10 and 80 articulatory and target dimensions, respectively, in our implementation). In this way, while group theoretical approaches offer a promising research line, they offer to date no general solution applicable to the multi-parametric speech control problem at hand.

The question of affine image registration, common in medical imaging applications among others, offers yet a different setting that can equally be characterized using Equation 2.8, and could potentially offer insight into possible solutions to the present problem. In fact rigid body transformations of an image follow equation 2.8, where now $\mathbf{x}$ is a vector representing the intensity of pixels in the image, the elements Λ_i represent the image translation and rotation parameters, and the matrices $\mathbf{H_i}$ characterize each of these individual actions. The affine image registration problem is, in this way, equivalent to the current inverse control problem (i.e. we wish to estimate the parameters Λ_i that would bring a present image closer to a second target image). A broad range of different methodologies has been proposed in the image registration literature to solve this problem. Brown (1992) offers a review of standard image registration techniques. One set of methods (feature- or label-based methods) involves the identification of a minimal subset of salient features existing in both the present and target images, which are then

coregistered using a least square regression or similar technique (e.g. Pelizzari et al. 1988) . In the current speech control context, these methods would resemble those involving formant extraction, as formants represent one of the most salient spectral features of speech sounds. Similarly as in the image coregistration problem, one drawback of formant-based methods are the limitations in our ability to uniquely define corresponding formants in both the present and the target sound, as the identity of each formant is not always obvious (e.g. formant crossings). A second set of common methods in the image registration literature (spatial- or pixel-based methods) attempt to minimize a global index of the difference between the two images. This minimization is usually performed using some variation of a gradient descent technique. For example, Woods et al. (1992) use an iterative Levenberg-Marquardt algorithm, a modification of the Gauss-Newton algorithm, on an intensity ratio measure of image discrepancy. Viola and Wells (1997) use a gradient descent on a mutual information measure of image discrepancy. The choice of discrepancy measure is usually problem-specific based on their sensitivity to expected sources of noise. The variations on the gradient descent algorithm usually offer increased robustness compared to a simpler gradient descent technique based on a least-square discrepancy measure. The reason is that the latter methodology tends to be affected by local minima, working best when the present and target images are relatively close to begin with. These examples are of closer relevance to the present problem of speech control. They suggest that a gradient descent technique could offer a viable solution for the present inverse control problem if appropriately modified to minimize the impact of local minima. In this section we will take this approach and directly extend the control equation 2.5, implementing a modified uni-variate gradient descent technique, to the multivariate case.

First, we will assume that we have a set of *N* anti-symmetric matrices $\mathbf{H}_i$ modeling the effect on the sound spectrum resulting from moving along *N* (possibly redundant) arbitrary directions in the articulatory space[7]. For any arbitrary sound spectra $\mathbf{x_1}$ and $\mathbf{x_2}$ we will define the combined measure $\mathbf{x_1 \Theta x_2}$ (which we will denote by exponential difference) as the *N*-dimensional vector with components determined by Equation 2.9a.

$$[\,\mathbf{x_1} \Theta \mathbf{x_2}\,]^{\mathbf{i}} = \mathbf{x}_1^{t} \cdot \delta(\mathbf{H}_i) \cdot \mathbf{x}_2 \qquad \textbf{Eq. 2.9a}$$

$$[\,\mathbf{x_1} \Theta \mathbf{x_2}\,]^{\mathbf{ij}} = x_1^i \cdot x_2^j - x_1^j \cdot x_2^i \qquad \textbf{Eq. 2.9b}$$

This measure has, in fact, some of the usual properties of a difference or subtraction of $\mathbf{x_1}$ minus $\mathbf{x_2}$. Specifically, $\mathbf{x_1} \Theta \mathbf{x_1}$ is always zero for any arbitrary vector $\mathbf{x_1}$, and $\mathbf{x_1} \Theta \mathbf{x_2}$ is always equal to $-(\mathbf{x_2} \Theta \mathbf{x_1})$, again for arbitrary vectors $\mathbf{x_1}$ and $\mathbf{x_2}$. Furthermore, for our problem at hand this measure covaries, as shown in the previous section, with the changes in the articulatory dimensions necessary to bring the sound spectrum $\mathbf{x_2}$ towards $\mathbf{x_1}$. This last property makes the exponential difference a desirable measure of the discrepancy between two sound spectra, especially for speech motor control.

Note that the exponential difference measure in Equation 2.9a depends on the definition of a set of matrices $\mathbf{H_i}$ characterizing the spectral transformations of interest. It is nevertheless possible to define an exponential difference measure that is independent of any set of generating matrices. This measure is shown in Equation 2.9b[8]. All exponential difference measures associated with

[7] We choose to define the system based on *N* arbitrary articulatory directions, rather than using an articulatory base (with minimal and independent dimensions), simply because it offers a more general approach.

[8] In this equation $\mathbf{x}^i$ represents the i-th element of the vector $\mathbf{x}$. The raw measure $\mathbf{x_1} \Theta \mathbf{x_2}$ in Equation 2.9b has been defined for simplicity as a matrix (using two indexes i and j). An equivalently definition

arbitrary sets of matrices $\mathbf{H_i}$ can be generated from this more generic measure by a simple linear transformation (see Appendix II.C). This result will be used in section III to define a raw acoustic measure of spectral difference (independent of the model matrices $\mathbf{H_i}$), from which a specific exponential difference measure (one depending on the model matrices $\mathbf{H_i}$) will be derived for the motor control of the speech articulators.

Again we wish to use an inverse control strategy now for the multi-parametric control of the articulators. The inverse control mechanism, extending the uni-parametric case in Equation 2.5, and now acting on the articulator descriptor **m**, will take the form:

$$\frac{d}{dt}\mathbf{m}(t) = \mathbf{G}\left(\mathbf{x}_{\text{target}}, \mathbf{x}(\lambda)\right) \qquad \textbf{Eq. 2.10a}$$

$$G\left(\mathbf{x}_{\text{target}}, \mathbf{x}(\lambda)\right) = \mathbf{F}\left(\mathbf{x_{target}} \ominus \mathbf{x(m)}\right) \qquad \textbf{Eq. 2.10b}$$

where $\mathbf{x_{target}}$ is the spectral target, $\mathbf{x}(\mathbf{m})$ is the current sound spectrum, and **F** is a vector function taking the exponential difference measure into a corrective motor command acting to reduce this difference. From the properties of the exponential difference it is reasonable to expect the function **F** to be well approximated by a linear function or a piece-wise linear function (this requires using also the information of the current articulatory configuration $\mathbf{m}(t)$ to appropriately account for the validity of the different exponential model approximations in the context of the current production). In other words, the function **F** relating differences in the spectrum to directional motor commands can now be approximated using a simple HRBF network. This is the

---

in terms of a vector difference measure is simply obtained using the vectorization operation (concatenation of the matrix columns).

same scenario as in the standard DIVA implementation when mapping differences in formant positions to directional motor commands.

The definition of the computation in Equation 2.9 (exponential difference) as a difference measure is not gratuitous. It is what brings the strategies proposed for the acoustic control of low dimensional parametric spaces such as formant dimensions into accordance with the control of the sound spectral shape. What the derivations in these sections indicate is that, if we substitute the usual measure of difference between sounds (in the case of formants, this would be the difference in formant positions) by the newly proposed exponential difference measure, we should be able to keep the usual strategies and assumptions for the control of the articulators (in this case, a simple HRBF network associating acoustic differences with changes in the articulator positions). This notion extends naturally to accommodate other measures of interest. For example, during an initial babbling phase, the mapping **F** is assumed to be learned by comparing the directions of change of the articulators to the direction of change in the sounds. Using the concept of exponential difference we can construct something equivalent to an "exponential derivative" that would measure the rate and direction of change in the sound spectrum. This would simply take the form:

$$\mathbf{dx}(t) \equiv \int \left(\mathbf{x(t+\tau)} \ominus \mathbf{x(t-\tau)} \right) \cdot \delta(\tau)\, d\tau$$

Eq. 2.11

where $\delta(\tau)$ is again (but unrelated to the previous usage in Equations 2.5 and 2.9a) the derivative of the *sinc* function. This equation is defined such that if we substitute the exponential difference

by a standard difference (a subtraction) the resulting measure would equal the standard derivative of **x** with respect to time.

Another example of a derived measure would be the definition of distance between two sound spectra (or relatedly the similarity between two spectra). A simple way of defining the distance between two sound spectra would be as the norm of the exponential difference between the two sound spectra. This definition again is such that if we substitute the exponential difference by a standard difference (a substraction) the resulting measure would equal the standard spectral distance (the norm of the difference of their log-spectra).

Before showing examples of the proposed multi-parametric control strategy acting on the vocal tract articulators to produce specific spectral targets, we need to deal with two issues. The first is the definition of the matrices $\mathbf{H}_i$ (i=1…N). These matrices are meant to model common spectral changes associated with articulatory movements. The second and inter-related issue is how this proposed strategy relates to known properties of auditory processing in the human brain. We would like to see if the computations involved in the proposed control strategies can be put in a simple form amenable to those commonly attributed to neural computations. In this way we would like to hypothesize a model of the neural computations involving the representation of and comparisons between speech sounds. Also we might be interested in learning from neurophysiology and behavioral studies what spectral changes the auditory system pays special attention to, thus aiding the construction of the yet-unspecified modeling parameters $\mathbf{H}_i$. This is the subject of the next section.

III. AUDITORY PLANNING AND NEURAL COMPUTATIONS

A. Spectral shape representation in auditory cortex

Axons of the auditory nerve project to the ventral and dorsal cochlear nuclei of the brainstem. From these regions new axons travel upwards to the superior olive and the inferior colliculus, projecting then to the medial geniculate body of the thalamus. From there auditory fibers project to the primary auditory cortex, a small region in the temporal lobe located inside the Sylvian fissure on the Heschl gyri (Brodmann areas 41 and 42). The response properties of cells in primary auditory cortex show an orderly spatial organization to the sound frequency content. This is known as tonotopic or cochleotopic organization, and it has been demonstrated in several mammalian species (Merzenich and Brugge, 1973, Merzenich et al., 1975; Reale and Imig, 1980; Schreiner and Cyander, 1984; Morel and Kass, 1992; Morel et al., 1993; Recanzone et al., 1999) as well as in humans (Le et al., 2001; Romani et al., 1982; Wessinger et al., 1997). Most neurons in primary auditory cortex seem to respond strongly to single frequency-modulated sweeps, and most of these responses are selective to the sweep direction and/or its rate (Brechmann et al., 2002; Hall et al., 2002; Mendelson and Cynader. 1985; Mendelson et al, 1993; Shamma et al., 1993, Zhang et al., 2003). These studies also present evidence of topographically organized responses to the stimulus sweep velocity in primary auditory cortex. Overall this literature seems to indicate that primary auditory cortex analyzes the sound spectrum locally in frequency, while placing special emphasis on local frequency modulations.

B. Auditory cortex and the DIVA model

While the accurate description of neural activation is extremely complex, many neural models drastically simplify its functional definition, emphasizing two main aspects. One is the linear nature of the combinations of multiple inputs; the second is the application of non-linear

transformations at the level of each neuron's activation. That is, the emphasis on these models is the ability to describe complex computations as a combination of linear operations on multiple elements, together with non-linear functions of individual elements, where these core elements are identified with the activation of individual neurons. We are going to take the same approach and propose a neural implementation of the spectral-target inverse control computations as a composition of linear combinations of multiple elements and element-wise nonlinear operations. The individual elements will be associated with the activation of auditory and motor cortical neurons. This approach is schematized in Figure 2.7. A spectral representation (**x**) of the present sound is defined as the cochlear output associated with this sound. This representation can be stored in short term memory and can be brought back again for the comparison between this reference sound and the sound currently being presented. A non-linear hard-coded operation initially compares the present sound with the one stored in short term memory to produce a raw measure (**z)** of their differences. Based on auditory experience, aimed at modeling natural spectral transformations in speech sounds, a higher-order area computes a measure of the spectral difference (DV) between these two sounds, corresponding to the exponential difference measure in our previous mathematical definitions, and equivalent to the Difference Vector in the current DIVA model. This computation is implemented by a set of linear weights. While beyond the goal of the current work, it is expected that simple learning rules could be defined to characterize these weights from auditory experience alone, as they model the spectral transitions occurring in natural speech sounds. In the present work we will explicitly define these weights to characterize local frequency shifts in the sound spectra (see following sub-section). Last, a piece-wise linear transformation (implemented with a RBF network) relates this spectral difference measure to the corrective motor command (DM) acting to reduce the discrepancy between the present and target sounds. As in the current DIVA model this transformation could be learned following simple

associate learning rules. In the present work we will explicitly define this transformation based on sample DV/DM pairs obtained during an initial babbling phase.

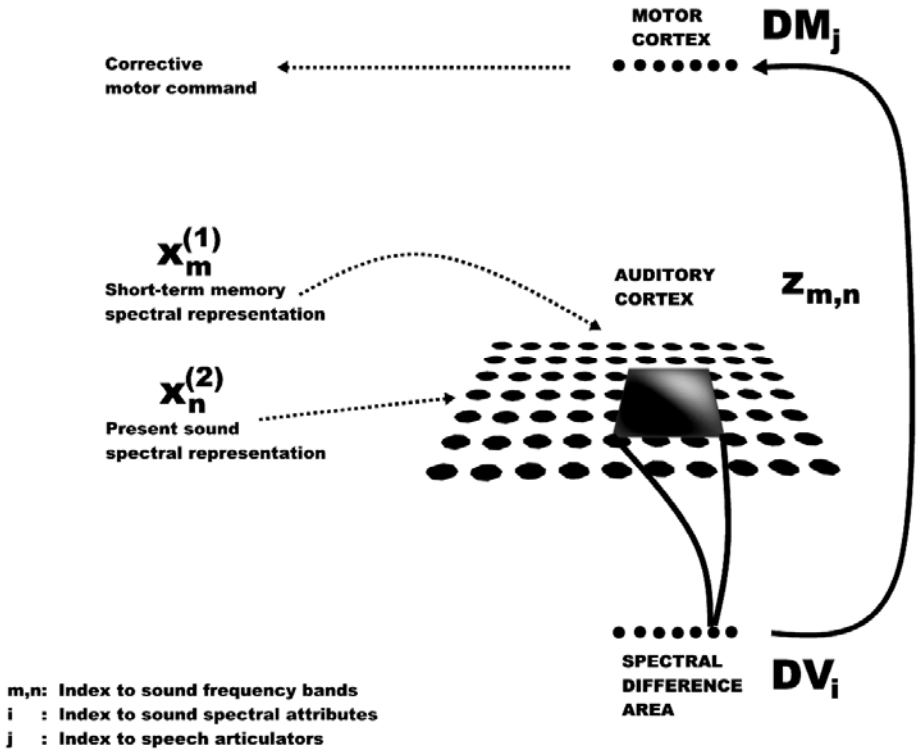

Figure 2.7 Schematic of the neural model for the implementation of the spectral target inverse control strategy. The cochlear representation of a sound ($\mathbf{x}^{(1)}$) stored in short-term memory, and a present sound ($\mathbf{x}^{(2)}$) are combined through a non-linear operation at the level of individual neurons to produce a raw measure of their spectral differences (**z**). Based on the auditory experience, a linear combination relates this measure with a spectral difference measure (**DV**) that focuses on spectral attributes characterizing natural spectral transitions in speech sounds (corresponding to the spectral difference measure in the previous sections). The **DV** measure is mapped through a piece-wise linear map onto a corrective motor command (**DM**) acting to reduce this discrepancy.

There are three elements involved in this computation. The first element is the nonlinear operation at the level of the cortical neurons performing a raw comparison of the spectral representations associated with two sounds. This is defined by the simple product:

$$z_{m,n} = \left[x_m^{(1)} \cdot x_n^{(2)} - x_n^{(1)} \cdot x_m^{(2)} \right]^+$$

Eq. 2.12

Where $\mathbf{x}^{(1)}$ is the cochlear representation stored in short term memory and $\mathbf{x}^{(2)}$ is the cochlear representation due to the current sound. The notation $[\ ]^+$ represents the rectification of the activation $z_{m,n}$ to positive only values. The resulting measure **z** stores a raw comparison between these two sounds. This raw measure corresponds in our previous discussion with the exponential difference measure that is independent of the definition of the matrices **H** (Equation 2.9b). If the sound stored in short term memory represents the sound presented in the recent past (i.e. it acts roughly as a short-time delay) then the measure **z** characterizes the dynamic aspects of the input sound **x**(t). If the sound stored in short term memory represent the target sound, the measure **z** characterizes the spectral difference between the present and target sounds. The second element is the set of linear weights $\mathbf{w}^{(A)}$ relating this to the spectral difference measure (Equation 2.9a) through an equation of the form:

$$DV_i = \sum_{m,n} w_{i,m,n}^{(A)} \cdot z_{m,n}$$

Eq. 2.13

The matrices of weights $\mathbf{w}_i^{(A)}$ correspond in our previous discussion to the modeling matrices $\delta(\mathbf{H}_i)$ (i.e. $w_{i,m,n}^{(A)} = [\delta(\mathbf{H}_i)]_{m,n}$). This definition make the **DV** measure exactly equal to our definition of the exponential difference $\mathbf{x}^{(1)} \Theta \mathbf{x}^{(2)}$ (Equation 2.9a) between the sound stored in short term memory and the currently present sound. In the neural model these weights take the form of local receptive fields on the neurons $z_{m,n}$, Conceptually they characterize the main features of interest in the activations $z_{m,n}$ resulting from speech sounds. While it is expected that these weights could be learned from auditory experience, in the present implementation we will define them explicitly to characterize local frequency shifts (see following sections).

In our emphasis on the dynamic aspects of the spectral shape, we have disregarded so far changes in amplitude. Now it is time to incorporate these aspects in the context of the current neural model. Conceptually we wish to incorporate into the activations **z** (computing a raw measure of spectral difference) a set of neurons that encode amplitude differences between the spectra. The simplest way to accomplish this is by adding a bias term in the spectral representation **x**. That is, we add a single element, set arbitrarily to a constant value of 1, to our original definition of the vector **x**. If **x** is a vector with L elements (80 frequency bands in our implementation) we will denote this element by its index $L+1$. Applying now Equation 2.12 to the extended vector **x**, the raw difference measure **z** incorporates a new set of neurons effectively measuring the amplitude changes in the spectrum for each frequency band. This is explicitly shown in the following equations, which directly follow from 2.12:

$$z_{m,L+1} = \left[x_m^{(1)} - x_m^{(2)}\right]^+$$

$$z_{L+1,m} = \left[x_m^{(2)} - x_m^{(1)}\right]^+$$

Eq. 2.12b

Correspondingly, we add a new element to the **DV** measure of the system in Figure 2.7. This element receives its inputs from the newly defined elements $z_{m,L+1}$ and $z_{L+1,m}$. Its weights are set to a value of +1 and -1, respectively, performing the following operation:

$$DV_{N+1} = \sum_m z_{m,L+1} - z_{L+1,m}$$

Eq. 2.13b

This new element is defined to simply reflect the difference in amplitude between the sound stored in short term memory and the currently present sound.

The third and last element in Figure 2.7 corresponds to the mapping between acoustic and motor differences (**DV** and **DM**, respectively). This mapping corresponds, in our mathematical description of the inverse control strategy, to the function **F** in Equation 2.10b. This mapping is implemented through a set of local weights $\mathbf{w}^{(B)}$:

$$DM_j = \sum_i w_{j,i}^{(B)}(\mathbf{m}) \cdot DV_i$$

Eq. 2.14

where the vector **DM** defines a corrective motor command in the articulator space (10 dimensional articulatory descriptor vector in our implementation), and the weights $\mathbf{w}^{(B)}$ are defined locally depending on the present articulatory configuration **m**. This definition is equivalent to the current DIVA model implementation of this mapping through a RBF network. These weights are learned from experience during an initial babbling phase, when {**DV**,**DM**,**m**} samples are generated.

This finishes the definition of the proposed neural model implementing the inverse control strategy for spectral targets in the context of the DIVA model. Section III.C will explicitly define the set of weights $\mathbf{w}^{(A)}$ and it will compare some features of the proposed neural model to auditory cortex neurophysiology. In section IV, the weights $\mathbf{w}^{(B)}$ will be defined and implementation examples of the inverse control strategy will be analyzed.

C. Empirical and modeling approximations to the exponential difference measure.

While we propose that the neural model weights $\mathbf{w}^{(A)}$ could be learned from early auditory experience to reflect those spectral transformations commonly occurring in speech sounds, in the present work we opted for a simpler approach that directly models them to describe one of the most common spectral transformations in speech (local frequency shifts). This section starts by substantiating that local frequency shifts in effect describe some of the most common spectral changes in speech, and then deriving the model weights $\mathbf{w}^{(A)}$ that characterize these transformations.

A purely empirical approximation to the definition of the exponential difference measure based on articulatory movements would require the estimation of a set of matrices $\mathbf{H}_i$ characterizing the

spectral changes associated with random movements of the articulators. From them, the weights $\mathbf{w}^{(A)}$ of the proposed neural model could be directly defined. We estimated 500 of these matrices, each approximating Equation 2.3b for a fixed direction of articulatory movement around an arbitrary articulatory configuration. Examples of a few of these matrices are shown in Figure 2.8 top left. These matrices characterize the acoustic energy transfers occurring when the articulators move in a given direction. One of the most prominent features of these empirical matrices is in the near-diagonal elements. These elements characterize transfers of acoustic energy between nearby frequency bands. This is corroborated by a principal component analysis (PCA) of these 500 matrices. Figure 2.8 bottom left plots the first 8 components of the resulting PCA, which show an accentuated role of near-diagonal elements in the first components.

These near-diagonal, relatively smooth (changing slowly along the matrix diagonal), elements characterize energy transfers between nearby frequency bands, and they are characteristic of local frequency shifts. From the neurophysiological data discussed above, local shifts in the sound spectrum also represent a possible prominent feature in the cortical analysis of sounds. Motivated by this combination of partial but coinciding evidence, we propose then a modeling approximation to the matrices $\mathbf{H}_i$ based on a set of matrices characterizing local frequency shifts in the spectrum. We created 63 of these matrices; a few of them are shown in Figure 2.8 bottom right. These matrices could be used to reconstruct the sample empirical matrices. This reconstruction is shown in the top right plot of Figure 2.8. As this plot indicates, the choice of modeling matrices $\mathbf{H_i}$ emphasizes those aspects of the empirical data corresponding to local frequency shifts and de-emphasizes the rest. Specifically 41% of the variance of the empirical matrices $\mathbf{H_i}$ is explained by the modeling approximation.

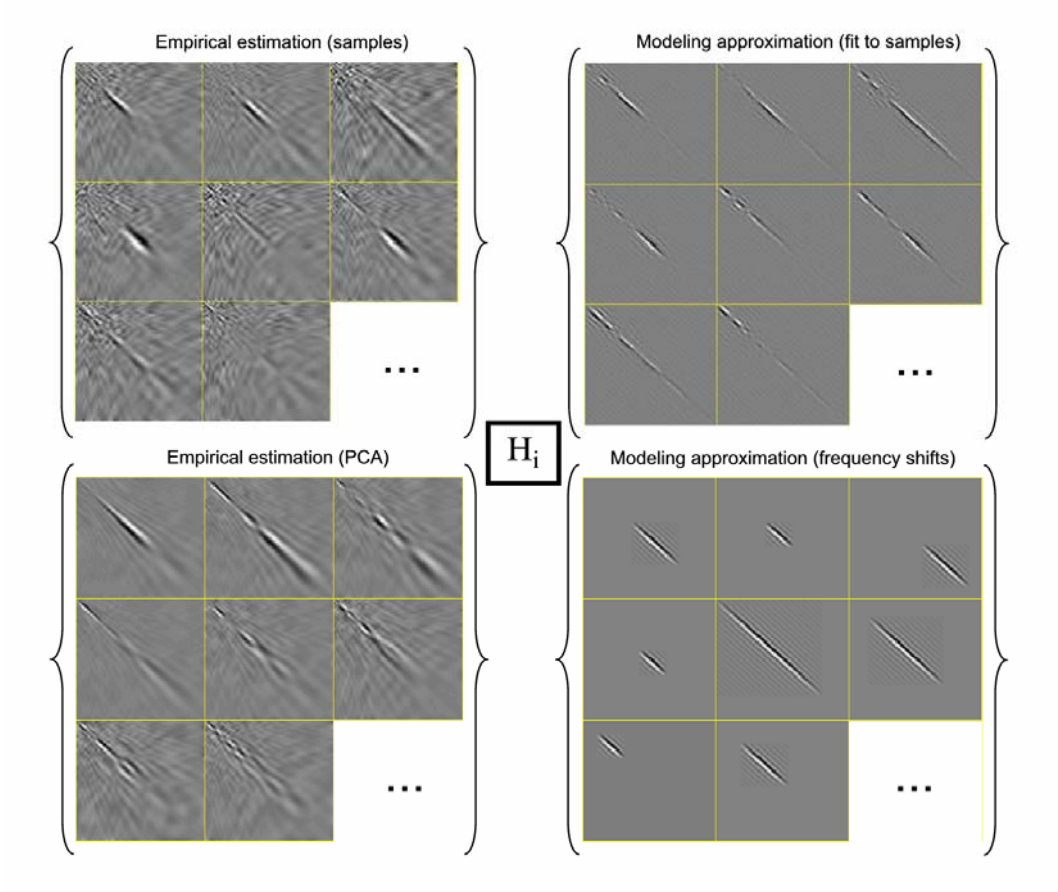

Figure 2.8 Estimation of model matrices $\mathbf{H}_i$ approximating the spectral changes associated with movements of the articulators. **Left**: Empirical estimation. Top plots shows 8 examples of model matrices resulting from fitting sample data of sound spectra associated with movement of the articulators along arbitrary articulatory directions. Each of these matrices can be interpreted as modeling the sound energy transfer between pairs of frequency bands, when the articulators move in a given direction. In these plots the horizontal and vertical axes represent the sound frequency. Intensity values code the rate of change of the sound energy at the frequency band characterized by the vertical axis, as a proportion of the sound energy at the frequency band characterized by the horizontal axis. The empirical matrices shown on the top were estimated using the algorithm in Appendix II.A. The bottom plots show the first 8 components from a principal component analysis (PCA) of these empirical $\mathbf{H}_i$ matrices. These plots, together with the sample plots above, highlight that the empirically estimated matrices $\mathbf{H}_i$ have a characteristic form with prominent

elements close to the diagonal and relatively smooth (changing slowly along the matrix diagonal). These elements characterize energy transfers corresponding to local frequency shifts of the sound spectra. **Right:** Modeling approximation. Matrices $\mathbf{H_i}$ can be generated from model matrices characterizing local frequency shifts with different center frequencies and extents of the shifts. Examples of 8 of these generating matrices are shown in the bottom plot. The top plot shows the reconstruction of the sample empirical matrices (those shown on the left top plot) using these generating matrices.

The choice of modeling matrices $\mathbf{H_i}$ was set to cover the frequency space with local frequency shifts characterized by a variety of center frequencies and ranges. Figure 2.9 top shows an example of these modeling matrices $\mathbf{H_i}$ (one that covers the whole frequency range) and its action on a random spectrum. This action can be interpreted as frequency shift saturating at the limits of the frequency range. The specific choice of center frequencies and ranges for the modeling matrices $\mathbf{H_i}$ is schematized in Figure 2.9. bottom. They form six scaling levels of frequency shift ranges, with 75% overlap of the frequency ranges on each level.

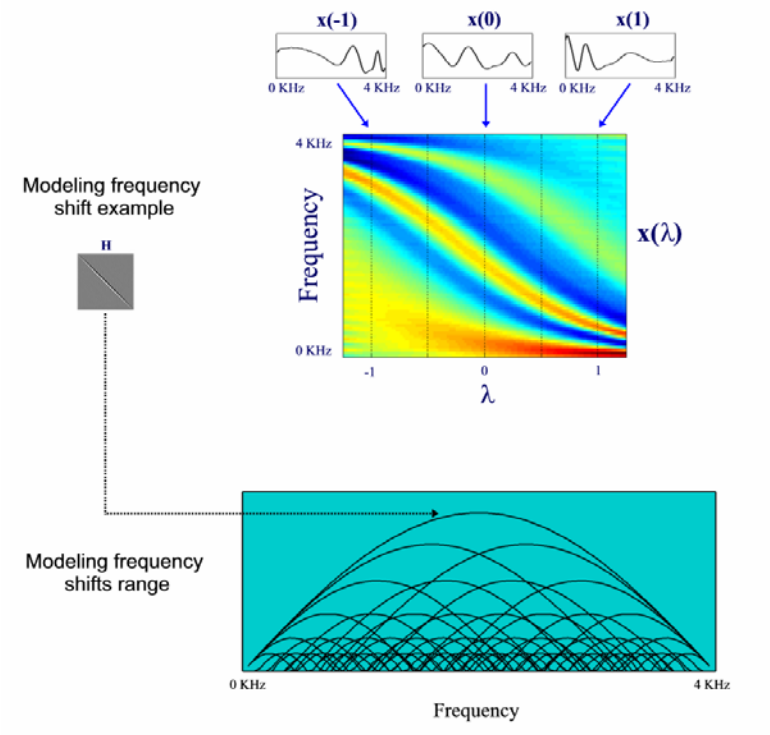

Figure 2.9 Modeling frequency shift matrices. **Top**: Example of the action of a model matrix **H** implementing a frequency shift over the whole sound spectrum. The direct application of the exponential model on this matrix produces a global frequency shift that saturates at the ends of the range of frequencies covered by the matrix. This effect is exemplified on a random sample spectrum. The model matrices $\mathbf{H_i}$ are defined as local versions of the one shown, covering different frequency ranges. **Bottom**: Schematics of the choice of center frequencies and ranges of the modeling matrices $\mathbf{H_i}$ approximating local frequency shifts. A total of 63 matrices represent frequency shifts with a variety of center frequencies and frequency ranges. There are 6 levels or "scales" of frequency shifts (represented by different heights in the plot). For each scale the number of frequency shifts is set to consecutive powers of two, and the frequency ranges are defined to have 75% overlap.

This definition of the modeling matrices $\mathbf{H_i}$ leads directly to an explicit definition of the neural model weights $\mathbf{w}^{(A)}$ used in our implementation (by defining $w_{i,m,n}{}^{(A)} = [\delta(\mathbf{H}_i)]_{m,n}$). An example of these weights is shown in Figure 2.7. As shown in this example the set of weights $\mathbf{w}^{(A)}_{\mathbf{i}}$ projecting

to a given element DV_i of the spectral difference measure takes generally the form of a bipolar receptive field on the cortical surface of the model neurons $\mathbf{z}_{m,n}$.

An example of the activations of the modeled neurons is shown in Figure 2.10. This example was constructed using the sounds /a/ and /e/, respectively, as the inputs for the short term memory and cochlear representations, respectively (for example, during the presentation of the sound /a-e/). The activation of the neurons $z_{m,n}$ forms a two dimensional image characterizing the differences between these sounds. In this image, high frequencies are represented in the bottom right corner, and low frequencies in the top-left corner. Neurons located in the lower triangular part respond to spectral differences that can be roughly characterized as energy shifting to the upper spectral frequencies. Neurons located in the upper triangular part respond to energy shifting to lower spectral frequencies. The spectral difference **DV** measure is composed by looking at the difference between the activation of upper and lower triangular segments with varying center frequencies and extents. This follows from the bipolar form of the matrices $\delta(\mathbf{H}_i)$ represent the weights between the neuron activation $z_{m,n}$ and the **DV** measure. The **DV** measure, in turn, analyzes the difference between these two sounds in terms of possible frequency shifts with different extents and center frequencies. The two most prominent features of the activations $z_{m,n}$ can be characterized as an upward frequency shift in the mid-high frequencies, and a downward frequency shift in the mid-low frequencies. These features are captured by the elements of the **DV** measure at different levels. Shown in the figure are the receptive fields of two modeled neurons capturing these aspects when focusing on a relatively large frequency scale. The elements DV_i focus on frequency shifts at a range of center frequencies and ranges or scales as defined above. The figure also illustrates the effect of the bias terms $z_{m,L+1}$ and $z_{L+1,m}$ (the right and bottom portions of the image, respectively). The activation of these neurons represents the difference in

spectral amplitude between the two sounds at different center frequencies. The corresponding last term of the **DV** measure computes the global amplitude change between these two sounds.

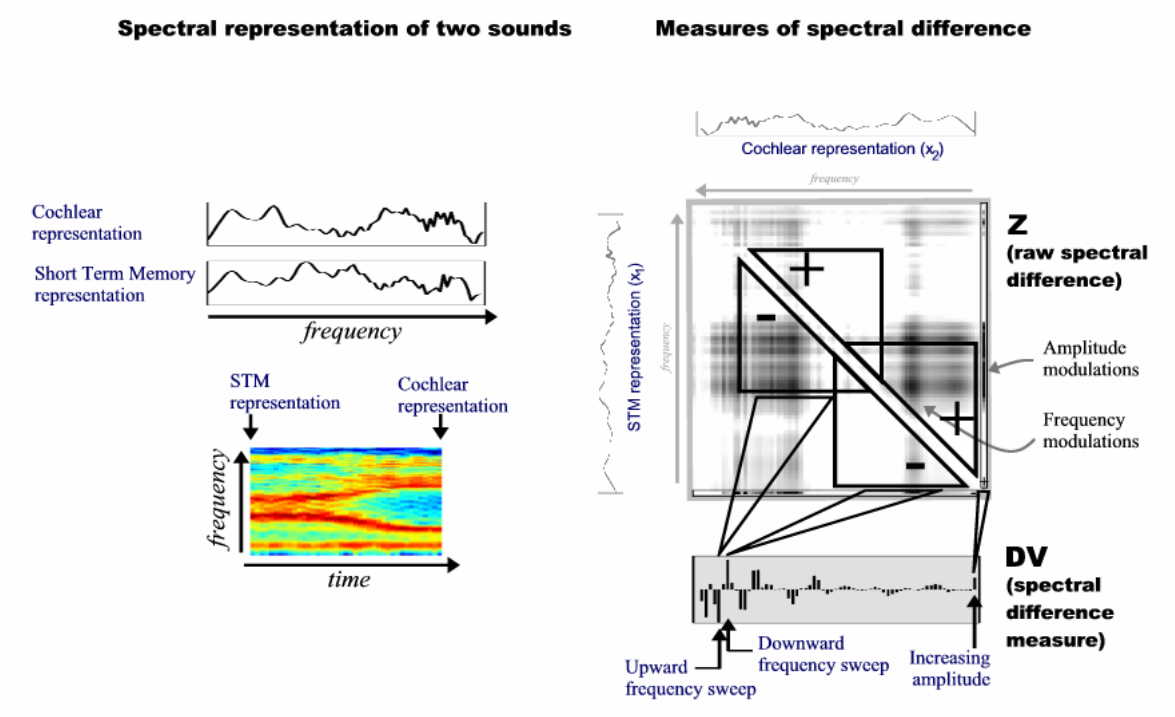

Figure 2.10 Example of modeled neuron's activations for a present sound /a/ and the memory trace of sound /e/. **Left:** The cochlear representation of the vowel /a/ and the short term memory copy of the sound /e/ are shown in the left plots. Also a natural transition between these sounds is shown in the bottom plot for reference. **Right:** The modeled neuron's activations $z_{m,n}$ (composing a raw spectral difference measure) are shown in the central figure. The **DV** spectral difference measure shown in the bottom looks at those specific features of the raw spectral difference characterizing local shifts in frequency. These are characterized by the **z** activations around its diagonal axis. Each element in the **DV** measure examines possible local frequency shifts with different extents and center frequencies between the two sounds. The last element of this measure, in contrast, examines differences in amplitude between the two sounds. Three example features are highlighted showing indications, respectively, of an upward frequency sweep in the mid-high frequencies, a downward frequency sweep in the mid-low frequencies, and an overall increase in amplitude.

In relating this model to neurophysiology or imaging data we will assume that, under normal circumstances, the short term memory trace is going to contain a representation of the sound presented in the recent past; in the case of inverse-control for speech the short term memory is assumed to have a representation of the target sound. Under this assumption the activation $z_{m,n}$ of modeled cortical neurons will respond strongly to frequency modulations, in particular upward (for m>n) or downward (for m<n) spectral peaks shifting between the m-th and n-th frequency bands during the time delay defining the short term memory representation. In contrast, the original activations x_n respond to the instantaneous frequency characteristics of the present sound (c.f. the activations $z_{m,n}$ representing its dynamic aspects). We will assume that the activations x_n are also present in auditory cortex. These activations follow the local energy of the present sound for the n-th frequency band, and they would be expected to show stronger activations for static sounds than dynamic ones. These observations so far would correspond with the general neurophyisiological evidence regarding tonotopical organization and directional sensitivity to frequency sweeps found in auditory cortical neurons. Last, the activations $z_{L+1,m}$ will respond strongly to sudden increases in the sound amplitude at the m-th frequency band, while the activations $z_{m,L+1}$ will respond to sudden decreases in amplitude. These would correspond with neurons coding amplitude modulations in the sound stimuli. These type of responses can also be found in studies of animal and human auditory cortex (Schreiner and Urbas, 1986; Hart et al. 2003). Regarding the short term memory trace we make no explicit assumptions about its cortical implementation. It could form a separate representation of the recent past with back projections to the modeled neurons $z_{m,n}$, or it could be implicitly realized in the way neurons $z_{m,n}$ respond to present vs. recent-past stimuli (dynamic neurons).

This description indicates that the modeled representations **x** and **z** can provide a first order approximation to the response properties and functional organization of auditory cortical neurons. The previous analyses also show that a secondary region receiving its projections from these neurons through a set of simply defined local weights can compute the exponential difference measure designed for the control of the speech articulators. In the following section we will show examples of the application of this model for the inverse control of speech using auditory spectral targets.

IV. EXAMPLES OF AUDITORY PLANNING USING A VOCAL TRACT SYNTHESIZER

In this section we will show examples of the proposed multi-parametric control strategy acting on the vocal tract articulators to produce specific spectral targets. In these examples the vocal tract model that we choose to control is the Maeda articulatory synthesizer (1990). As discussed in previous section, the inverse-control strategy for speech involves a mapping (weights $\mathbf{w^{(B)}}$ in the neural model) between the movement of the articulators and the associated spectral differences. A locally linear approximation was hypothesized to suffice. We estimated this mapping explicitly from sample data. The sample data was generated using 10,000 productions of short linear articulatory movements. For each production we stored the trio {$\mathbf{m_n}$, $\mathbf{dm_n}$, $\mathbf{dv_n}$ }where $\mathbf{m_n}$ is the average (center) articulatory configuration of the production, $\mathbf{dm_n}$ is the direction of articulatory movement (or average articulatory change), and $\mathbf{dv_n}$ is the average spectral change. The mapping $\mathbf{w^{(B)}}$ was explicitly estimated at each iteration of the inverse-control strategy from the sample data {$\mathbf{m_n}$, $\mathbf{dm_n}$, $\mathbf{dv_n}$ } for the current articulatory configuration **m** as:

$$\mathbf{w}^{(\mathrm{B})}(\mathbf{m}) = \left[\sum_n e^{-\sigma^{-1}\left\|\mathbf{m}-\mathbf{m}_n\right\|^2} \cdot \mathbf{dm}_n \cdot \mathbf{dm}_n^t \right]^{-1} \cdot \sum_n e^{-\sigma^{-1}\left\|\mathbf{m}-\mathbf{m}_n\right\|^2} \cdot \mathbf{dm}_n \cdot \mathbf{dv}_n^t$$

The parameter σ is set to 0.1 (it affects how many sample data are deemed close enough to the current articulatory configuration **m** to be considered for the local estimation of $\mathbf{w^{(B)}}$).

The initial simulation (Figure 2.11) exemplifies the proposed inverse control strategy. The controller was acting on the articulatory parameters to approximate a spectral target defined from the articulatory synthesizer to approximate the American English vowel sound /a/. The initial

state (n=1 in Figure 2.11) was set to a neutral configuration of the articulators with zero glottal pressure (for silence). From this configuration the inverse controller used equations 2.12-2.14 to define incremental changes to the articulators through 100 iterations. In each iteration the exponential difference measure **DV** (equivalently $\mathbf{x}_1 \Theta \mathbf{x}_2$) between the target spectrum and that of the current production was computed, and the mapping $\mathbf{w}^{(B)}$ was used to transform the spectral difference into an incremental motor command. In order to bias the system towards relaxed configurations and to speed up the convergence, two small factors were added to the incremental motor command: a relaxation factor (driving the articulators towards a centered configuration), and an inertia factor (temporally smoothing the articulatory trajectory) [9].

In this simulation (Figure 2.11) from the initial to approximately the fifth iteration the main change in the articulators corresponds to an increase in sound amplitude, which is due to an increase of the glottal pressure. This mainly acts to decrease the originally large amplitude error. From this to approximately the 20th iteration the system acts to correct the error in the spectral shape of the produced sound to better approximate the target spectrum. This is accomplished mainly by an opening of the mouth together with a backwards movement of the tongue. After this, the system continues incrementally fitting more detailed aspects of the spectrum (limited by the inherent noise in the sound analysis/synthesis process). The final configuration produces a sound with a spectrum mimicking the desired target spectrum (2% mean square error of the cochlear representation of the sound spectrum).

[9] These factors are added to the incremental motor command at each iteration. The relaxation factor consists of a term proportional to the difference between a fixed (average) articulatory configuration and the present articulatory configuration. The inertia factor consists of a term proportional to the incremental motor command at the previous iteration.

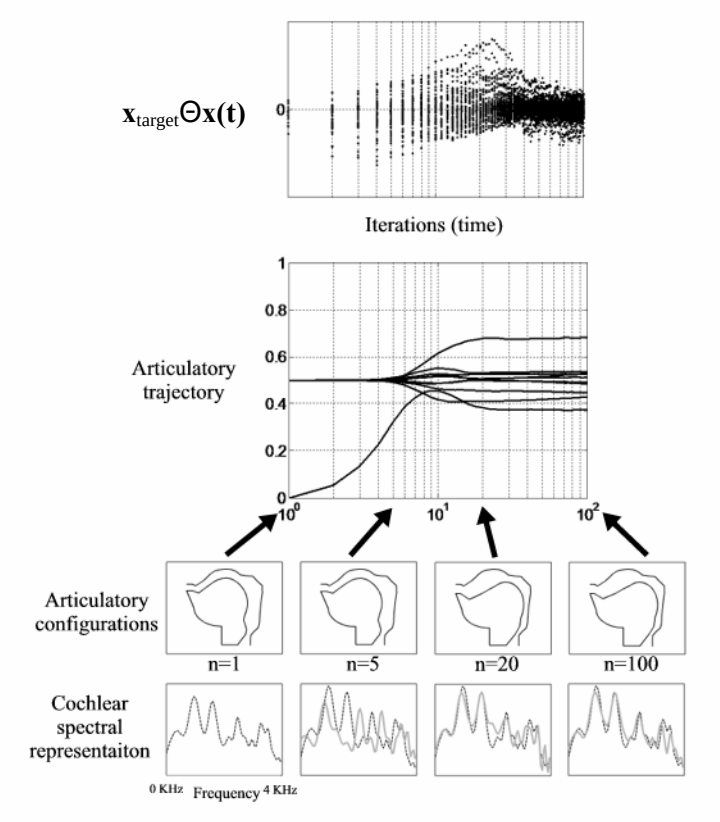

Figure 2.11 Time-line of inverse control of the vocal tract articulators using spectral targets. The target spectrum approximates that of the /a/ vowel production (shown in dashed line in the bottom plots). The top plot shows the measure of spectral difference between the target sound and the system's production at each iteration of the inverse control loop. The middle plot shows the resulting trajectory of the vocal tract articulatory parameters. Four relevant moments of this trajectory are highlighted in the bottom plots, showing their corresponding vocal tract configurations and produced spectra.

The next simulation extends the previous example by using target spectra defined by productions of the articulatory synthesizer roughly approximating standard vowel sounds for American English. The results (Figure 2.12) show a good approximation of the target spectra (between 2% and 8% mean square error when comparing the final and target cochlear representations), as well as the formants of the target spectrum (between 1% and 4% error for the first two formants,

which are considered important cues in the perceptual identification of vowels). The configurations that the articulators adopt also mimic standard configurations for the corresponding vowels. In particular the final configurations comply with the front-back and high-low phonetic distinctions among vowels (in Figure 2.12 the productions from left to right can be categorized as front-middle-back, and high-low-high).

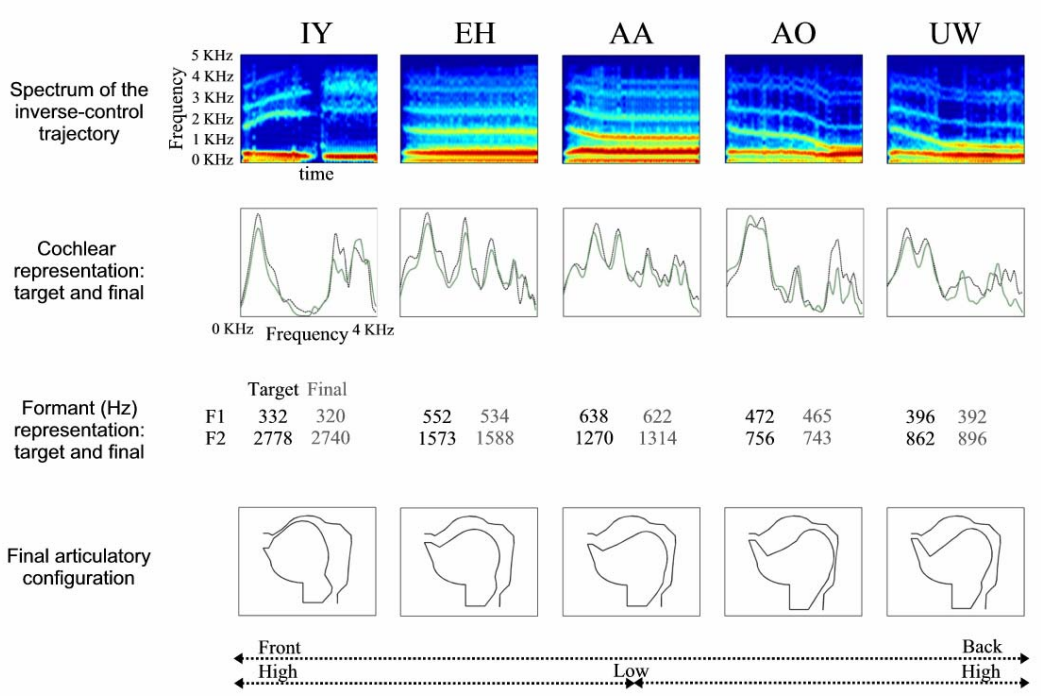

Figure 2.12 Simulations of the inverse controller acting on an articulatory synthesizer to mimic static spectral targets. Top row shows the dynamic sound spectrum of the inverse control trajectories (starting from a silent rest configuration). The target spectra were defined by productions of the articulatory synthesizer characterizing typical vowel sounds. The final configurations approximate the desired spectral target, as indicated by a good fit between the cochlear as well as the formant representations of the target and final sounds. Furthermore the final articulatory configurations mimic those of typical speakers, matching the front-back and high-low phonetic distinctions between vowel sounds.

Last we addressed the system's robustness to the definition of spectral targets based not on the own productions of the articulatory synthesizer, but on a human speaker's vowel productions (Figure 2.13). In order to emphasize both typical vowels and glide sounds, we used in this example a single simulation where the target was changed every 100 iterations. The spectral targets cycled through the five Spanish vowel sounds (the author's native language). Due to the differences in voice and filtering between the human speaker and the articulatory synthesizer, the target spectra are typically non-reachable targets. As expected, the error in the system final productions is larger in these examples (18% and 10% mean error in the first two formants, respectively). Nevertheless, the produced sounds are perceptually identifiable as the target phonemes, and the system produces natural transitions between the vowel sounds. Interestingly only limited aspects of the target spectra are mimicked. While the mean square error of the target spectrum is quite large (between 30% and 60% mean square error), the features that make these sounds recognizable (the formants, or relative positions of spectral peaks) are still present in the system's productions. This relative robustness of the system stems from the local analysis of the spectrum. This is similar to other techniques that de-emphasize the large scale spectral characteristics of the sound (e.g. liftering, in cepstral analysis), a commonly used strategy to reduce the sensitivity to variations in speaker characteristics, vocal efforts, variations in transmission, etc. (see Rabiner and Juang, 1993).

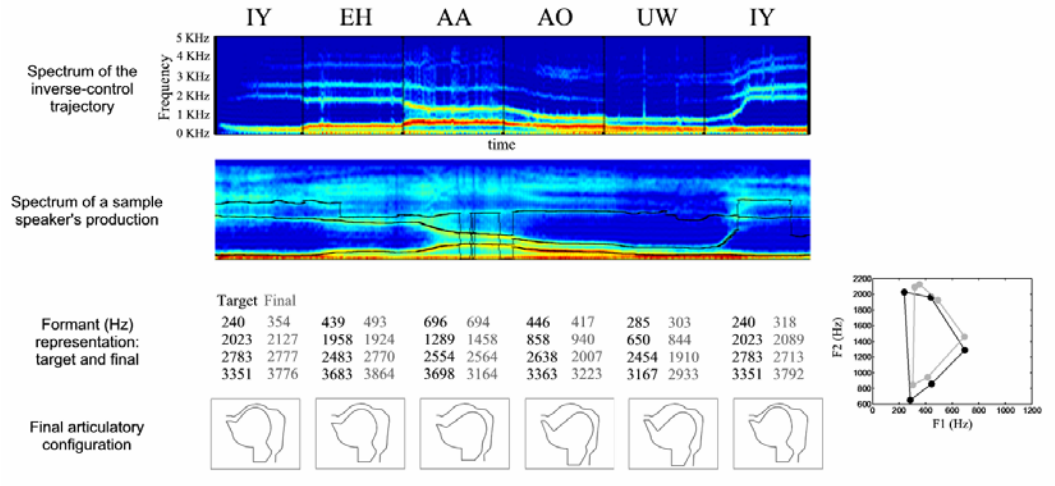

Figure 2.13 Simulations of the inverse controller acting on an articulatory synthesizer to mimic static spectral targets. The target spectra were defined by static vowel productions of a Spanish speaker. Top row shows the dynamic sound spectrum of the inverse control trajectories (the targets are changed every 100 iterations of the inverse controller). For reference the second row shows the production of the same vowel sequence by the original speaker with superimposed tracking of the first three formant positions. The bottom rows show the final articulatory configuration for each of the six vowels, together with the formant representation of the produced sound at this point and the formants of the target vowel spectra. The target and final configurations of the first two formant positions, which are important cues for the perceptual recognition of vowels, are shown on the right.

The previous examples show that the proposed multi-parametric control strategy can be used to operate on the speech articulators to produce acoustic targets defined by their spectral characteristics. Before these techniques can be effectively applied to produce the variety of sounds in natural speech there are many questions that need to be addressed. One is the relevance of dynamic spectral information in speech control. A simple inverse control technique can produce stop consonants simply by tracking the spectrum leading into and out of the closure. This is demonstrated in Figure 2.14, where the sequences /aba/ and /aga/ are produced using the same

inverse control strategy as in the examples above, now using dynamic spectral targets (that is, the spectral target is redefined for each iteration). In this example the system approximately reproduces the correct place of articulation for the consonant productions.

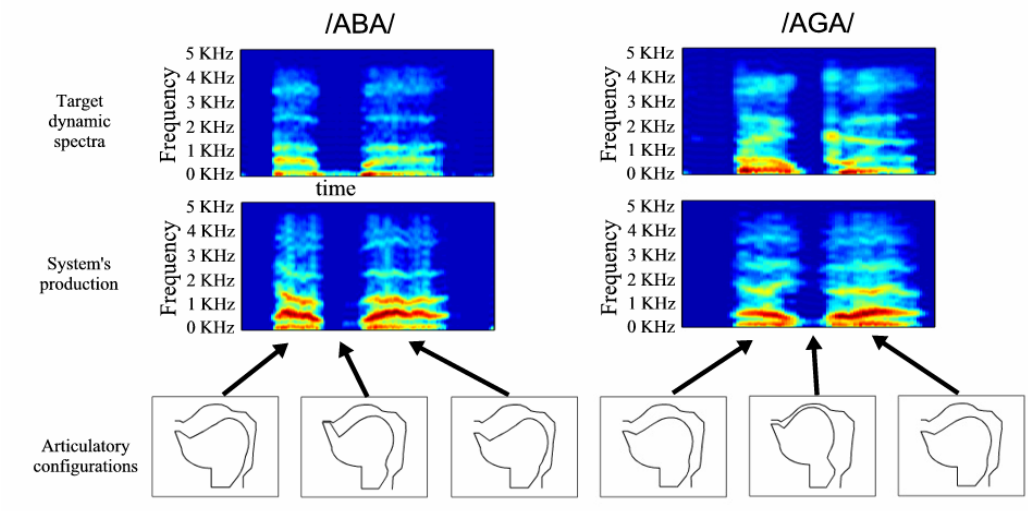

Figure 2.14 Simulations of the inverse controller acting on an articulatory synthesizer to mimic dynamic spectral targets. The target spectra were defined by a speaker's productions of the sequences /aba/ (left) and /aga/ (right). The inverse-controlled system mimics the spectral targets, and uses appropriate places of articulation for the stop consonant productions.

Nevertheless a proportional inverse control is in general going to have limited ability in effectively tracking dynamic spectra. This is mainly due to its disregard for the dynamic aspects of the sound it is trying to mimic. A derivative term (as in a proportional-derivative controller) could potentially improve the performance for dynamic control for speech. While the acoustic measures necessary for such control are already available using the proposed exponential model (the exponential derivative concept introduced in section II.B) appropriate implementation of these techniques is beyond the goals of this thesis. Furthermore, the optimization of the inverse

controller for the case of dynamical targets might be unnecessary. The presence of a learned feedforward command in the DIVA model is expected to iteratively reduce the necessary corrections from the dynamical inverse controller, alleviating the necessity to optimize the latter.

Another issue is the question of speaker independence. We have not yet dealt with one important source of speaker variability in the spectral characteristics of their productions, and that is the size of their vocal tract. This is particularly important for children learning to speak, who have a considerably smaller vocal tract than many of the speakers that form their acoustic influence. This variability mainly results in a speaker-specific scaling of the vocal tract resonances. In terms of the cochlear representation (close to logarithmic-frequency scaled) this translates approximately into a global spectral shift. One strategy to deal with this, and possibly other, sources of undesired variability would be to remove them through pre-processing. In this case, for example, one could use the same inverse-control strategies to affect not only the articulators but also a pre-processing stage where the cochlear spectrum would undergo a global shift. This approach would nevertheless need validation. Another approach would be to change the original spectral representation to be invariant to global shifts of the spectrum (for example, the absolute value of the cepstral coefficients). Yet it is unclear at this point what strategy the human auditory system uses, or could potentially use, to deal with this issue.

Another practical issue relates to the sensitivity to specific acoustic differences. The model proposed in this chapter takes the general form of local frequency shifts based partly on descriptions of the neurophysiology of auditory cortical neurons, and also partly based on the characterization of common acoustic changes associated with speech articulatory movements. Human sensitivity to different features of speech sounds is nevertheless non-uniform, and it

seems to be affected by language-dependent auditory experience. In the proposed model, the sensitivity to different acoustic features is mainly related to their relative importance in the definition of the spectral difference measure. It would be desirable to be able to bias the spectral difference measure towards increased sensitivity to specific spectral features. In some cases this could be accomplished by the definition of new matrices $\mathbf{H_i}$ (or in terms of the cortical model, the weights $\mathbf{w^{(A)}}$) that would typify the desired features. In the more general case we could use empirically defined matrices $\mathbf{H_i}$ (such as those shown in Figure 2.8) characterizing typical spectral trajectories in the base language instead of, or on top of, the currently defined ones. Our original simulations (not shown) indicate that efficient control of the articulators is equally possible using empirically defined matrices instead of the currently used modeled ones. This opens the possibility of extending the model through empirical estimation procedures from sample acoustic data while incorporating the effects of experience. Last, there are also a number of acoustic features that call for an extension of the current cochlear representation, rather than the cortical model. For example, information about the sound voicing and pitch are currently not present (or at least not salient) in the cochlear representation due to the combination of implicit (small time windows) and explicit (cepstral low-pass) smoothing of the spectrum.

Last, the strategies presented for inverse control are based on an error correction algorithm. While this has been presented in the context of inverse online control of the speech articulators it should be also possible to phrase them in terms of error correction learning algorithms (both of these strategies are present in the DIVA model, in the feedback and feedforward control loops, respectiely). It is reasonable to expect, under an auditory target hypothesis, that children learning to mimic the spectro-temporal features of phonetic or syllabic units common in their language use some form of error correction to incrementally improve their productions over the learning

period. The inverse control strategies presented in this chapter could potentially be used to define learning rules for the incremental motor specification of well practiced articulatory trajectories (feedforward learning rules). In either case the questions regarding timing and the incorporation of intrinsic delays need further analysis.

V. CONCLUSIONS

This chapter presented the application of inverse control strategies on the speech articulators using auditory spectral targets. This work fills the gap between the experimental results in the first chapter indicating a speech motor control system which effectively uses auditory formant targets, the experimental indications of multifaceted sp vectral sound representations in auditory cortex, and the desire to extend the definition of auditory targets to sounds which are not possible to characterize by their formants positions. The proposed exponential model approximates the relationship between the articulators and the associated acoustics, when the speech articulators are defined based on effective articulatory dimensions and the acoustics of the sounds are defined based on a biologically based cochlear sound representation. The exponential model approximation is used to propose control strategies tailored to the estimated articulatory-acoustic relationship. Throughout this work special emphasis has been put in creating models that are well-defined, yet extendable and simple to interpret. Towards this end the definition of the exponential difference measure (a novel measure of the acoustic difference between two sounds) plays a crucial role. It allows the simple integration of the mathematical complexities of the studied models into the context of previous efforts using formant sound representations. The general mathematical models defined in this chapter have been concretized to reflect central aspects of the cortical representation of sounds. Using these models the results presented in this chapter emphasize the possibility of effective control of the speech articulators using spectral auditory targets.

APPENDIX

APPENDIX I.A. Derivation of articulatory/acoustic relation from the motor control equations of the DIVA model.

In the DIVA model, the differential equation governing the articulator vector **x***(t)* given an acoustic target vector ***y*** takes the form:

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{J}^{+}\cdot\left(\mathbf{y}-\mathbf{f}\left(\mathbf{x}(t)\right)\right)-\alpha\cdot\mathbf{\Pi}\left(\mathbf{J}\right)\cdot\mathbf{x}(t)$$

where **f**(**x**) represents the articulatory to acoustic mapping, **J** represents the Jacobian (the multivariate derivative) of this mapping at each point **x**(t), $\mathbf{J}^{+}$ and $\mathbf{\Pi}(\mathbf{J})$ represent its pseudoinverse and its null space projector operator, respectively, and α is a small factor in the model (relaxation factor) controlling the degree of articulatory relaxation toward a neutral configuration (without loss of generality this is assumed to be **x**=0). Under a linear approximation of the articulatory to acoustic mapping (**f**(**x**)=**A**·**x**), and using a regularized form of the pseudoinverse, the explicit form of the previous equation is:

$$\begin{aligned}\frac{d}{dt}\mathbf{x}(t) &= \mathbf{A}^{t}\cdot\left(\mathbf{A}\cdot\mathbf{A}^{t}+\mu\cdot\mathbf{I}\right)^{-1}\cdot\left(\mathbf{y}-\mathbf{A}\cdot\mathbf{x}(t)\right)-\alpha\cdot\left[\mathbf{I}-\mathbf{A}^{t}\cdot\left(\mathbf{A}\cdot\mathbf{A}^{t}+\mu\cdot\mathbf{I}\right)^{-1}\cdot\mathbf{A}\right]\cdot\mathbf{x}(t)\\ &= \mathbf{A}^{t}\cdot\left(\mathbf{A}\cdot\mathbf{A}^{t}+\mu\cdot\mathbf{I}\right)^{-1}\cdot\left(\mathbf{y}-\left(1-\alpha\right)\cdot\mathbf{A}\cdot\mathbf{x}(t)\right)-\alpha\cdot\mathbf{x}(t)\end{aligned}$$

where **A** is the linear mapping between the articulatory and acoustic spaces, and μ is a small regularization factor of the pseudoinverse. The solution of this differential equation is the articulatory trajectory **x***(t)*:

$$\mathbf{x}(t) = \mathbf{x}_0 + \left(\mathbf{I}-e^{-\mathbf{K}\cdot t}\right)\cdot\left(\mathbf{x}_\infty-\mathbf{x}_0\right)$$
$$\mathbf{K} \equiv \left(1-\alpha\right)\cdot\mathbf{A}^{t}\cdot\left(\mathbf{A}\cdot\mathbf{A}^{t}+\mu\cdot\mathbf{I}\right)^{-1}\cdot\mathbf{A}+\alpha\cdot\mathbf{I}$$

where $\mathbf{x_0}$ is the initial articulatory configuration, and $\mathbf{x_\infty}$ is the articulatory configuration that would be reached allowing infinite time ($\mathbf{x_\infty}$ depends on the acoustic target **y**, and its solution is not relevant to the following discussion). Repeated productions under different initial articulatory configurations, will reach, after time *T*, the articulatory configuration **x***(T)*, following a distribution with average:

$$\langle \mathbf{x}(T) \rangle = \mathbf{x}_\infty - e^{-\mathbf{K}\cdot T} \cdot \left(\mathbf{x}_\infty - \langle \mathbf{x}_0 \rangle \right)$$

and covariance:

$$\mathbf{\Omega}_T = e^{-\mathbf{K}\cdot T} \cdot \mathbf{\Omega}_0 \cdot e^{-\mathbf{K}^t \cdot T}$$

where $\langle \mathbf{x_0} \rangle$ and $\mathbf{\Omega_0}$ are the average and covariance, respectively, of the initial articulatory configurations. For simplicity, let us assume the distribution of initial articulatory configurations to be normal, with covariance $\sigma_0 \cdot \mathbf{I}$. In this case, the articulatory covariance of the final articulatory configurations takes the form:

$$\mathbf{\Omega}_T = \sigma_0 \cdot e^{-2\cdot\mathbf{K}\cdot T}$$

Let us, finally, define the vector **q** to be any eigenvector of the matrix $\mathbf{\Omega}_T$ (corresponding with one of the articulatory directions resulting from a principal component analysis of the final articulatory covariance). The *acoustic effect* of this articulatory direction **q** is defined as the associated change in the acoustic vector when moving the articulators along the direction **q**, and it is computed as $\lambda(\mathbf{q}) \equiv \left\| \mathbf{A} \cdot \mathbf{q} \right\|$, and the *articulatory variability* associated with the same articulatory direction **q** is computed as $\sigma(\mathbf{q}) \equiv \mathbf{q}^t \cdot \mathbf{\Omega}_T \cdot \mathbf{q}$. Using the definition of the matrices $\mathbf{\Omega}_T$ and **K**, and noting that their eigenvectors (they are the same for both matrices) will correspond to the right- eigenvectors of the matrix **A**, the articulatory variability $\sigma(\mathbf{q})$ can be expressed, as a function of the acoustic effect $\lambda(\mathbf{q})$, as:

$$\sigma(\mathbf{q}) = \sigma_0 \cdot e^{-2\cdot\left[(1-\alpha)\cdot\frac{\lambda^2(\mathbf{q})}{\lambda^2(\mathbf{q})+\mu}+\alpha\right]\cdot T}$$

More simply, the articulatory/acoustic relation predicted from the DIVA equations belongs to the class of functions:

$$\sigma(\lambda) \propto \varepsilon^{\frac{\lambda^2}{\lambda^2+\mu}}$$

where ε and μ are two small factors. The dashed line in Figure 1.8 left is an example of such a function approximating the simulation results (ε=.08; μ=.03).

APPENDIX II.A. Solution of symmetry-constrained linear equations

We want to solve a linear equation of the form:

$$\mathbf{y} = \mathbf{A} \cdot \mathbf{x}$$

where **y** and **x** are two arbitrary data matrices, and **A** is the unknown, which is further constrained to be anti-symmetric:

$$\mathbf{A} + \mathbf{A}^T = 0$$

The constrained least-square solution will obey (using the Euler-Lagrange multiplier theorem):

$$\mathbf{y}\mathbf{x}^T - \mathbf{A}\mathbf{x}\mathbf{x}^T - \mathbf{B} = 0$$

where **B** is a symmetric matrix. The solution of the original equation is, then:

$$\mathbf{A} = \left(\mathbf{y}\mathbf{x}^T - \mathbf{B}\right) \cdot \left(\mathbf{x}\mathbf{x}^T\right)^{-1}$$

where the term $\mathbf{y}\mathbf{x}^T\left(\mathbf{x}\mathbf{x}^T\right)^{-1}$ corresponds to the unconstrained solution of the original linear equation, which we will denote as $\mathbf{A}_0$. The matrix **B** must be solved to satisfy the original symmetry-constraint on **A**, which now takes the form:

$$\mathbf{B} \cdot \left(\mathbf{x}\mathbf{x}^T\right)^{-1} + \left(\mathbf{x}\mathbf{x}^T\right)^{-1} \cdot \mathbf{B} = \mathbf{A}_0 + \mathbf{A}_0{}^T$$

The general solution of this last equation is (see Lyapunov's equation, Horn et al. 1991):

$$vec(\mathbf{B}) = \left(I \otimes \left(\mathbf{x}\mathbf{x}^T\right)^{-1} + \left(\mathbf{x}\mathbf{x}^T\right)^{-1} \otimes I\right)^{-1} \cdot vec\left(\mathbf{A}_0 + \mathbf{A}_0{}^T\right)$$

where *vec* represents the vectorization operation (concatenation of the columns of a matrix), *I* is the identity matrix, and ⊗ is the Kronecker product. The implementation of this solution is far simpler to compute using the reduced form:

$$\mathbf{B} = \mathbf{U} \cdot \left[\left(\mathbf{U}^{\mathbf{T}} \mathbf{D} \mathbf{U} \right) \circ \mathbf{E} \right] \cdot \mathbf{U}^{T}$$
$$\mathbf{D} \equiv \mathbf{A}_0 + \mathbf{A}_0^{\,T}$$
$$e_{ij} \equiv \left(s_i^{-1} + s_j^{-1} \right)^{-1}$$

where **U**, and **s** are the eigenvectors and eigenvalues, respectively, of $\mathbf{x}\mathbf{x}^{\mathrm{T}}$ (i.e. $\mathbf{x}\mathbf{x}^{T} = \mathbf{U} \cdot diag(\mathbf{s}) \cdot \mathbf{U}^{T}$), and "$\circ$" represents the Hadamard product (entrywise or Schur product). This reduced form was obtained using the definition of the eigenvalues and eigenvectors of the Kronecker product of two matrices. The solution of **A** is, then:

$$\mathbf{A} = \mathbf{A}_0 - \mathbf{U} \cdot \left[\left(\mathbf{U}^{\mathbf{T}} \mathbf{D} \mathbf{U} \right) \circ \mathbf{F} \right] \cdot \mathbf{U}^{T}$$

where the elements of the new matrix **F** are defined as $f_{ij} \equiv e_{ij} \cdot s_j^{-1}$. Finally, if the original constraint was for **A** to be symmetric (instead of anti-symmetric), the solution would be computed simply changing the definition of the matrix **D** to be:

$$\mathbf{D} \equiv \mathbf{A}_0 - \mathbf{A}_0^{\,T}$$

APPENDIX II.B. Exponential model inverse approximation

Let $\mathbf{x_1}$ and $\mathbf{x_2}$ be two real valued positive vectors with norm one, and $\mathbf{H}$ be a real-valued anti-symmetric matrix. We would like to find a scalar λ between -1 and 1, such that $e^{\mathbf{H}\cdot\lambda}\cdot\mathbf{x}_2$ best approximates $\mathbf{x_1}$ (in a least-squares sense). This optimization problem can be stated as:

$$\hat{\lambda} = \arg\min_{\lambda\in[-1,1]}\left\{\left|\mathbf{x}_1 - e^{\mathbf{H}\cdot\lambda}\cdot\mathbf{x}_2\right|^2\right\} = \arg\max_{\lambda\in[-1,1]}\{p(\lambda)\}$$

$$p(\lambda) \equiv \mathbf{x}_1^t \cdot e^{\mathbf{H}\cdot\lambda}\cdot\mathbf{x}_2$$

While this is a difficult non-linear problem, a useful approximation can be defined by limiting the range of searched λ values and treating $p(\lambda)$ as a probability density[10]. In this way, the previous equation defines λ as the *mode*, or peak, of the distribution $p(\lambda)$. Instead of its mode, we will settle to estimate its *expected value*, or average. In this way, the estimation of λ becomes:

$$\hat{\lambda} = \int_{-1}^{1} \lambda \cdot p(\lambda)\, d\lambda$$

Substituting the definition of $p(\lambda)$ we obtain:

$$\hat{\lambda} = \int \lambda \cdot \mathbf{x}_1^t \cdot e^{\mathbf{H}\cdot\lambda}\cdot\mathbf{x}_2^t \; d\lambda = \mathbf{x}_1^t \cdot \int \lambda \cdot e^{\mathbf{H}\cdot\lambda}\; d\lambda \cdot \mathbf{x}_2^t = \mathbf{x}_1^t \cdot \delta(H)\cdot\mathbf{x}_2^t$$

$$\text{where } \delta(z) \equiv \int \lambda \cdot e^{z\cdot\lambda} d\lambda = \frac{d}{dz}\int e^{z\cdot\lambda} d\lambda = \frac{d}{dz}\,\text{sinc}\left(\frac{z}{j\pi}\right)$$

Figure 2.15 left shows the form of the function $\delta(z)$. Note that this function, as defined, takes an imaginary number as input z and returns also a purely imaginary number $\delta(z)$. Also this function is anti-symmetric (i.e. $\delta(-z)=-\delta(z)$). In terms of matrix functions (see Appendix II.A), $\delta(\mathbf{H})$ transforms a real-valued anti-symmetric matrix into another real-valued anti-symmetric matrix.

---

[10] While $p(\lambda)$ does not comply with the definition of a probability density function (i.e. it is always positive, integrating to one) we find the relaxation of these constraints to lead to a simpler estimation that is similarly effective in our simulations.

As a last note, the limitation of the range of λ over the extent -1 to 1 does not imply any limitation on the scale of the search, as the matrix **H** is unconstrained (i.e. if we wanted to search over the range -2 to 2, we would simply multiply **H** by 2). A possible extension would be to include an a-priori function $q(\lambda)$ in the integration over λ instead of hard-limiting the range of the variable λ. This would result in a definition of δ(z) based on the derivative of the z-transform (closely related to the Fourier transform) of the function $q(\lambda)$. Figure 2.15 right shows an example of the resulting function $\delta(z)$ when $q(\lambda)$ is defined as a hamming window. The exact form would of course depend on the length of the hamming window used.

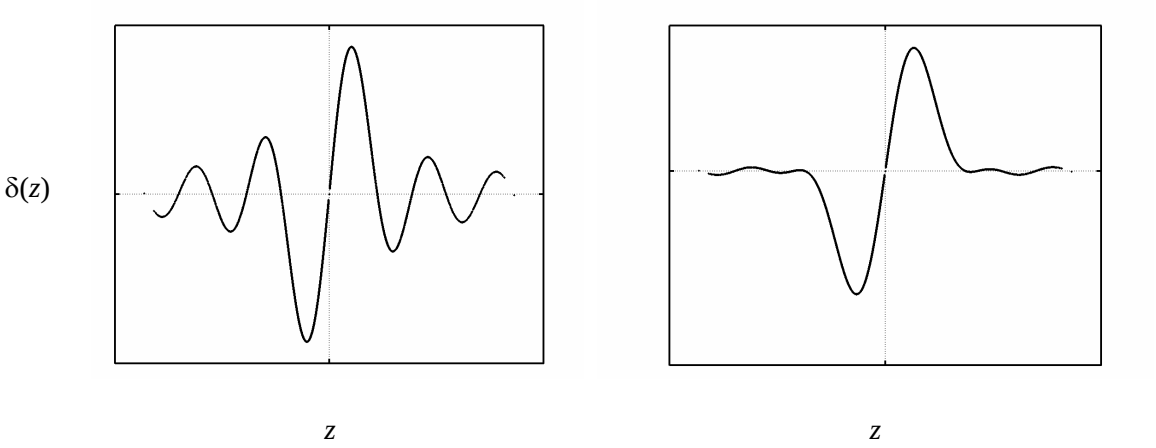

Figure 2. 15 Form of the function $\delta(z)$. The function $\delta(z)$ represents the relation between the eigenvalues of **H** (x-axis) and the weights of the exponential inverse approximation (y-axis). **Left:** The search range of λ is defined between -1 and 1. The function $\delta(z)$ can be characterized as the derivative of a *sinc* function. **Right:** Introducing a short hamming window a-priori $q(\lambda)$ and extending the search range to cover all the real line bounds and smoothes the resulting $\delta(z)$ function.

APPENDIX II.C. A short reference to concepts in matrix analysis

If a square matrix of arbitrary dimensions **A** has elements a_{ij}, the transpose of **A** (denoted by $\mathbf{A}^t$) is defined as the matrix with elements a_{ji}^* (the conjugate of a_{ji}). A matrix **A** is called diagonal if its only non-zero elements are located along the diagonal of the matrix ($a_{ij} = 0$, if $i \neq j$). It is called symmetric if its transpose is equal to the original matrix ($\mathbf{A}^t = \mathbf{A}$), and anti-symmetric if its transpose is equal to minus the original matrix ($\mathbf{A}^t = -\mathbf{A}$). It is called unitary if $\mathbf{A}^t \cdot \mathbf{A}$ and $\mathbf{A} \cdot \mathbf{A}^t$ equal the identity matrix (a diagonal matrix with ones in its diagonal), and normal simply if $\mathbf{A}^t \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A}^t$. All diagonal, symmetric, anti-symmetric or unitary matrices are normal matrices.

Any normal matrix **A** can be composed as the product $\mathbf{Q} \cdot \mathbf{D} \cdot \mathbf{Q}^t$ of a unitary matrix **Q** and a diagonal matrix **D**. The elements in the diagonal of the matrix **D** are called the eigenvalues of **A**, and the columns of the matrix **Q** are called the eigenvectors of **A**. Even when all the elements of the matrix **A** are real, the matrices **Q** and **D** are generally complex valued. In fact the eigenvalues of any anti-symmetric matrix are always purely imaginary. The eigenvalues of a symmetric matrix are always real, and those of a unitary matrix lie in the unit circle (they have absolute value equal to one). These relations also hold in the opposite direction (i.e. a matrix with purely imaginary eigenvalues is always anti-symmetric, etc.) Furthermore if **A** is a symmetric matrix, $j \cdot \mathbf{A}$ is anti-symmetric, and vice versa; and for any arbitrary matrix **A**, the matrix $\mathbf{A}+\mathbf{A}^t$ is always symmetric, and $\mathbf{A}-\mathbf{A}^t$ is always anti-symmetric. A simple rule-of-thumb to remember many of these matrix relations is to associate normal matrices with complex numbers, symmetric matrices with purely real numbers, anti-symmetric matrices with purely imaginary numbers, and unitary matrices with unit-norm complex numbers. Following this rule of thumb one would expect, for example, that we would get a unitary matrix by exponentiating an anti-symmetric matrix. We will

now see how this is in fact the case, but first we have to properly define what a function of a matrix is.

Powers of a matrix $\mathbf{A}^n$ are defined in the natural way as the product of the matrix **A** *n*-times. A new matrix $f(\mathbf{A})$ defined as a function of an original matrix **A**, where the scalar function $f(z)$ admits a Taylor decomposition $f(z)=r_0+r_1 \cdot z+r_2 \cdot z^2+r_3 \cdot z^3+\ldots$, is defined from the powers of the original matrix **A** as $f(\mathbf{A})=r_0+r_1 \cdot \mathbf{A}+r_2 \cdot \mathbf{A}^2+r_3 \cdot \mathbf{A}^3+\ldots$ If the matrix **A** admits a decomposition as $\mathbf{Q} \cdot \mathbf{D} \cdot \mathbf{Q}^t$ then the matrix $f(\mathbf{A})$ can be expressed as $\mathbf{Q} \cdot f(\mathbf{D}) \cdot \mathbf{Q}^t$. Furthermore, the matrix $f(\mathbf{D})$ is simply the diagonal matrix with elements $f(d_{ii})$. In other words, the eigenvectors of $f(\mathbf{A})$ are the same as the eigenvectors of **A**, and if d_{ii} are the eigenvalues of **A**, then the eigenvalues of $f(\mathbf{A})$ are $f(d_{ii})$. The exponentiation of a matrix is the function $f(\mathbf{A})$ where *f* is the exponential function (with a Taylor series $r_n=1/n!$) and it is denoted by $e^{\mathbf{A}}$. Two useful notions regarding matrix exponentiation are the following. First, the exponential of an anti-symmetric matrix is always unitary. This follows from the eigenvalues of $e^{\mathbf{A}}$ being the exponential of purely complex numbers (the eigenvalues of A). Second, the exponential of the sum of two matrices $e^{\mathbf{A}+\mathbf{B}}$ is in general different than $e^{\mathbf{A}} \cdot e^{\mathbf{B}}$ which in turn is different than $e^{\mathbf{B}} \cdot e^{\mathbf{A}}$. These three matrices are equal only if $\mathbf{A} \cdot \mathbf{B}=\mathbf{B} \cdot \mathbf{A}$ (then it is said that matrices **A** and **B** commute). Note that two arbitrary matrices generally will not commute (i.e. $\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}$), but a matrix always commutes with itself so $e^{\mathbf{A}(\lambda 1+\lambda 2)}$ always equals $e^{\mathbf{A}\lambda 1} \cdot e^{\mathbf{A}\lambda 2}$.

A most useful definition in matrix analysis is the notion of a matrix form, also called a bi-linear form. The form of a matrix **A** is defined as the following function:

$$f_A(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^t \cdot \mathbf{A} \cdot \mathbf{x}_2$$

where $\mathbf{x_1}$ and $\mathbf{x_2}$ are two arbitrary vectors. If the matrix **A** and the vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ are real-valued, the form of the matrix **A** is also always real. Bi-linear forms of symmetric matrices play an important role as they are commonly used to derive different notions of distance between two vectors $\mathbf{x_1}$ and $\mathbf{x_2}$. The form associated with an anti-symmetric matrix **A** can be directly related to the form associated with the symmetric matrix $j \cdot \mathbf{A}$ as $f_{\mathbf{A}}(\mathbf{x}_1,\mathbf{x}_2) = -j \cdot f_{j \cdot \mathbf{A}}(\mathbf{x}_1,\mathbf{x}_2)$. We will use the notion of bi-linear forms associated with real-valued anti-symmetric matrices in this dissertation as the basis for comparing two sounds. As a last observation, the bilinear form of any matrix **A** can always be expressed as a product of two vectors: one independent of **A**, and another depending on it, in the following way:

$$f_A(\mathbf{x}_1, \mathbf{x}_2) = vec(\mathbf{x}_1 \cdot \mathbf{x}_2^t)^t \cdot vec(\mathbf{A})$$

where *vec* represents the vectorization operation that takes a matrix and converts it into a vector by concatenating the matrix' columns. Thus all bilinear forms can be expressed as a linear combination of the product $\mathbf{x}_1 \cdot \mathbf{x}_2^t$ (this is called the outer product of the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$).

REFERENCES

Blackburn, C.S. (1996) “Articulatory methods for speech production and recognition” PhD thesis. University of Cambridge.

Boyce, S., and Espy-Wilson, C.Y. (1997). “Coarticulatory stability in American English /r/,” Journal of the Acoustic Society of America, 101, 3741-3753.

Brechmann A, Baumgart F, Scheich H. (2002). “Sound-level-dependent representation of frequency modulations in human auditory cortex: a low-noise fMRI study.” Journal of Neurophysiology, 87(1), 423-33.

Brown, L.G. (1992). “A survey of image registration techniques,” ACM Computing Surveys. 24(4), 325-376

Carrozzo, M., Stratta, F., McIntyre, J., and Lacquaniti, F. (2002), “Cognitive allocentric representations of visual space shape pointing errors,” Experimental Brain Research, 174(4), 426-436.

Delattre, P., Freeman, D.C. (1968). “A dialect study of American r’s by x-ray motion picture,” Lingusitics 44, 29-68.

Espy-Wilson, C.Y., and Boyce, S.E. (1994). “Acoustic differences between “bunched” and “retroflex” variants of American English /r/,” Journal of the Acoustic Society of America, 95, 2823.

Fant, G. (1980). “The relations between area functions and the acoustic signal,” Phonetica, 55-86.

Feldman, A. G. (1966) “Functional tuning of the nervous system with control of movement or maintenance of a steady posture. II. Controllable parameters of the muscles” Biophysics 11, 565-78.

Fowler, CA. (1990). “Sound-producing sources as objects of perception: rate normalization and nonspeech perception,” Journal of the Acoustic Society of America, 88(3), 1236-49.

Greenwood, D.D. (1961a). “Auditory masking and the critical band.” Journal of the Acoustic Society of America, 33, 484-502.

Greenwood, D.D. (1961b). “Critical bandwidth and the frequency coordinates of the basilar membrane.” Journal of the Acoustic Society of America, 33, 1344-1356.

Guenther, F.H., Espy-Wilson, C.Y., Boyce, S.E., Matthies, M.L., Zandipour, M., Perkell, J.S. (1999). “Articulatory tradeoffs reduce acoustic variability during American English /r/ production.” Journal of the Acoustic Society of America, 105(5), 2854-65.

Guenther, F.H., Ghosh, S.S., and Nieto-Castanon, A. (2003), “A neural model of speech production,” Proceedings of the 6th International Seminar on Speech Production, Sydney, Australia.

Guenther, F.H., Hampson, M., and Johnson, D. (1998). “A theoretical investigation of reference frames for the planning of speech movements,” Psychological Review 105, 611-633.

Hall DA, Johnsrude IS, Haggard MP, Palmer AR, Akeroyd MA, Summerfield AQ. (2002). “Spectral and temporal processing in human auditory cortex.” Cerebral Cortex, 12(2), 140-9.

Hart HC, Palmer AR, Hall DA. (2003). “Amplitude and frequency-modulated stimuli activate common regions of human auditory cortex.” Cerebral Cortex, 13(7), 773-81

Le, TH, Patel S, and Roberts, TP. (2001). “Functional MRI of human auditory cortex using block and event-related designs.” Magnetic Resonance in Medicine, 45, 254-60.

Maeda, S. (1990). “Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model,” in Speech Production and

Speech modeling, edited by W.J. Hardcastle and A. Marchal (Kluwer Academic, Boston), 131-149.

McIntyre, J., Stratta, F., Droulez, J., and Lacquaniti, F. (2000). “Analysis of pointing errors reveals properties of data representations and coordinate transformations within the central nervous system,” Neural Computation, 2(12), 2823-55.

MacNeilage, P.F. (1970). “Motor control of serial ordering of speech,” Psychological Review, 77(3), 182-196

Mendelson JR, Schreiner CE, Sutter ML, Grasse KL. (1993). “Functional topography of cat primary auditory cortex: responses to frequency-modulated sweeps.” Experimental Brain Research, 94(1), 65-87.

Mendelson JR, Cynader MS. (1985). “Sensitivity of cat primary auditory cortex (AI) neurons to the direction and rate of frequency modulation.” Brain Research, 18:327(1-2):331-5

Merzenich MM, Brugge JF. (1973). “Representation of the cochlear partition of the superior temporal plane of the macaque monkey.” Brain Research, 50(2),275-96.

Merzenich MM, Knight PL, Roth GL. (1975). “Representation of cochlea within primary auditory cortex in the cat.” Journal of Neurophysiology, 38(2), 231-49.

Morel A, Kaas JH. (1992). “Subdivisions and connections of auditory cortex in owl monkeys.” Journal of Comparative Neurology. 318(1), 27-63.

Morel A, Garraghty PE, Kaas JH. (1993). “Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys.” Journal of Comparative Neurology, 335(3), 437-59.

Payan, Y., and Perrier, P. (1997). “Synthesis of V-V sequences with a 2D biomechanical tongue model controller by the equilibrium point hypothesis,” Speech Communations, 22, 185-205.

Pelizzari, C.A., Chen, G.T.Y., Spelbring, D.R., Weichselbaum, R.R. and Chen, C.T. (1988). "Accurate three-dimensional registration of CT, PET and MR images of the brain," Journal of Computer Assisted Tomography, 13, 20-6

Perkell, J.S., Nelson, W.L. (1985). "Variability in production of the vowels /i/ and /a/," Journal of the Acoustic Society of America, 77(5), 1889-1895.

Perkell, J.S., Matthies, M.L., Svirsky, M.A., Jordan, M.I. (1993). "Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot "motor equivalence" study," Journal of the Acoustic Society of America, 93(5), 2948-2961.

Perkell, J.S., Matthies, M.L., Svirsky, M.A., Jordan, M.I. (1995). "Goal-based speech motor control: a theoretical framework and some preliminary data," Journal of Phonetics. 23, 23-35.

Perrier, P., Boe, L.J., and Sock, R. (1992). "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: modeling the transition with two sets of coefficients," Speech and Hearing Research, 35(1), 53-67.

Reale RA, Imig TJ. (1980). "Tonotopic organization in auditory cortex of the cat." Journal of Comparative Neurology, 192(2), 265-91.

Recanzone GH, Schreiner CE, Sutter ML, Beitel RE, Merzenich MM. (1999). "Functional organization of spectral receptive fields in the primary auditory cortex of the owl monkey." Journal of Comparative Neurology, 27;415(4), 460-81.

Romani, G.L., Williamson, S.J., and Kaufman, L. (1982). "Tonotopic organization of the human auditory cortex." Science 216(4552), 1339-40.

Rubin, P.E., Baer, T., and Mermelstein, P. (1981). "An articulatory synthesizer for perceptual research," Journal of the Acoustic Society of America, 70, 321-328.

Saltzman, C., and Munhall, K.G. (1989). “A dynamical approach to gestural patterning in speech production,” Ecological Psychology, 1, 333-382.

Schreiner CE, Sutter ML. (1992). “Topography of excitatory bandwidth in cat primary auditory cortex: single-neuron versus multiple-neuron recordings.” Journal of Neurophysiology, 68(5), 1487-502.

Shcreiner CE, Urbas JV (1986). “Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF).” Hearing Research, 21, 227-241

Schroeder, M.R. (1967). “Determination of the geometry of the human vocal tract by acoustic measurements,” Journal of the Acoustic Society of America, 41, 1002-1010.

Shamma SA, Fleshman JW, Wiser PR, Versnel H. (1993). Organization of response areas in ferret primary auditory cortex. Journal of Neurophysioly, 69(2), 367-83.

Stevens, S.S., and Volkmann, J. (1940). “The relation of pitch to frequency: A revised scale.” American Journal of Pyschology, 53, 329-353.

Story, B.H., Titze, I.R., and Hoffman, E.A. (1996). “Vocal tract area functions from magnetic resonance imaging,” Journal of the Acoustic Society of America, 100(1), 537-554.

Story, B.H., Titze, I.R., and Hoffman, E.A. (1998). “Vocal tract area functions for an adult female speaker based on volumetric imaging,” Journal of the Acoustic Society of America, 104(1), 471-487.

Tiede, M.K., and Yehia, H. (1996). “A shape-based approach to vocal tract area function estimation,” Proceedings ASA-ASJ 3rd Joint Meeting, 861-866.

Viola, P., and Wells, W.M. (1997). “Alignment by maximization of mutual information,” International Journal of Computer Vision, 24(2), 137-154

Wakita, H. (1973). “Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms,”. IEEE Transactions on Audio and Electroacoustics, AU-21 (5), 417-427.

Wessinger, C.M., Buonocore, M.H., Kussmaul, C.L., and Mangun, G.R. (1997). “Tonotopy in Human Auditory Cortex Examined With Functional Magnetic Resonance Imaging.” Human Brain Mapping 5(1), 18-25.

Woods, R. P., Cherry, S. R., and Mazziotta, J. C. (1992). “Rapid automated algorithm for alignment and reslicing PET images,” Journal of Computer Assisted Tomography, 16(4), 634-639

Zhang LI, Tan AY, Schreiner CE, Merzenich MM. (2003). Topography and synaptic shaping of direction selectivity in primary auditory cortex. Nature. 424(6945), 201-5.

Zwicker, E., Flottorp, G., and Stevens, S.S. (1957). “Critical bandwidth in loudness summation.” Journal of the Acoustic Society of America, 29, 548-557.

VITA

Alfonso Nieto-Castanon was born in Gijón, Spain, in September 1972, received his University degree in Telecommunication Engineering in Valladolid in 1996, and his PhD in Cognitive and Neural Systems in Boston in 2004. Along the way he developed a predilection for reading Latin-American literature, playing classical music, eating Spanish food, sharing time with friends, and writing on coffee-shop napkins.