


Universität des Saarlandes
Computational Linguistics and Phonetics

LORIA/INRIA Speech Research

Erasmus Mundus European Masters in
Language and Communication Technologies

Fitting an articulatory model with EMA data: toward an evaluation of speech inversion

Thesis Submission for a

Master of Science in
Language Science and Technology

(Specialization in Phonetics and Speech Technology)

Defended by

Mathew Wilson

Supervisors:	*Submitted:*
William Barry	September 20, 2008
Ingmar Steiner	Saarbrücken, Germany

Erklärung

Ich erkläre an Eides statt, dass ich die Masterarbeit mit dem Titel *Fitting an articulatory model with EMA data: toward an evaluation of speech inversion* selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt, und alle den benutzten Quellen wörtlich oder sinngemäss entnommenen Stellen als solche kenntlich gemacht habe.

Cambridge (Ontario, Kanada), den 1. September 2008

Mathew Wilson

Abstract

Imagine a device that could listen to a speaker's utterance and—on the basis of the acoustic signal and without any visual input—display the shape of her vocal tract in real-time. Among other applications, a language learner could use this visual feedback to improve pronunciation. In such an application, for speech inversion to be successful, the accuracy of its acoustic-to-articulatory mappings is essential. Current evaluation techniques are limited because they draw from either generalized phonetic knowledge or older X-ray tracings for a different speaker. The need for improved evaluation techniques underlies the goal of the work presented here: to draw from speaker-specific articulatory data, with a focus on tongue shape and position. For acoustic input, we consider a series of vowel-vowel sequences, the tongue movements for which we collect by three-dimensional (3D) electromagnetic articulography (EMA) from an adult male speaker of French. To facilitate eventual comparison with predicted vocal tract shapes from speech inversion, we need to express these 3D EMA data in the same format as the inversion output: as a series of 2D articulatory model parameters. We achieve this through a series of linear algebra transformations, supplemented by geometric speaker adaptation and numerical analysis. Although variations in approach and a quantifiable measure of fit remain as areas of further research, a visual inspection of results suggests that we now have a reasonable fitting technique in place.

Contents

1 Introduction

A device that could listen to a speaker's utterance and—on the basis of the acoustic signal and without any visual input—could display the shape of her vocal tract in real-time would have widespread applications. A learner of French could use this visual feedback to improve vowel pronunciation in words like *tu* and *puce*, a notorious hurdle for many native speakers of English. Similarly, a patient undergoing remedial pronunciation training after a stroke could benefit from such a technology. Artists in the animation industry could automate efforts to synchronize the movement of visible speech articulators to an audio track. When the vocal tract configuration is expressed in the form of just several articulatory parameters, speech inversion could act as a form of low bit-rate encoding for a much more complex speech signal. The excitement surrounding these applications and others forms a large part of the motivation behind acoustic-to-articulatory speech inversion.

Researchers seeking to someday develop speech inversion to this level of promise face several challenges. Above all, inversion may not be possible for all sounds of speech—and at the very least is more difficult for certain sound classes [19]. Also, the mapping of acoustic properties of speech to vocal tract shapes is not unique, where hundreds—perhaps approaching an infinity—of different configurations produce the same speech signal [13]. Moreover, a nonlinear relationship exists between deformation of the vocal tract during speech and the amount of change in the acoustics: a small change in tongue position in one region might result in a large change in the speech sound, whereas a significant movement elsewhere might cause little or no change in acoustics at all [19].

Even if these inherent challenges are overcome, for speech inversion to be successful, the accuracy of its acoustic-to-articulatory mappings is essential. Robust evaluation techniques are currently underdeveloped and tend to rely on (1) agreement between original and modelled acoustic data, or (2) qualitative assessment of static vocal tract shapes [14]. This strong need for improved evaluation techniques underlies the goal of the work presented

here. Specifically, our aim is to contribute to the development of an evaluation technique that draws from natural articulatory data, focusing on the shape and position of the tongue. The scope of the acoustic input that we consider for speech inversion is a subset of French vowels, namely /**i**/, /**a**/ and /**u**/.

An important stage in the direction of this work, but beyond its current scope, is the direct comparison of inversion results to a corpus of real-world vocal tract shapes registered over time—using the same speaker and the same acoustic signal. With this eventual target in mind, we drew from a corpus of recently collected speech articulator movements in three-dimensional (3D) space, recorded by electromagnetic articulography (EMA) from an adult male speaker of French.

To facilitate such an evaluation, we wish to represent these EMA data in the same currency of output used by the existing speech inversion technique at LORIA (Lorraine Institute for Computer Science and its Applications) in Nancy, France. There, inversion uses a complex acoustic-to-articulatory lookup table (a hypercube codebook) and a series of algorithms that reduce the many possible shapes to a single best solution. This solution is expressed in terms of articulatory parameters, based on the same two-dimensional (2D) articulatory model that was used to build the codebook. Therefore, the overarching goal of this project is to convert a set of EMA sensor positions in 3D space to a corresponding set of model parameters.

Following our work, a general technique for fitting EMA tongue data to Maeda's articulatory model is now in place, although variations of certain steps remain areas of further consideration. Examining the inclusion of mouth parameters will be important, since we expect a better fit from the additional data, and because we should take advantage of the 3D EMA data we have collected from the mouth fleshpoints. At this point the approach reviewed here is ready to proceed with preliminary comparisons to the results of speech inversion—but awaits multiple refinements in method and tools before being comfortably suitable for frequent and meaningful evaluation efforts.

2 Background

To optimally review a sufficient background for this work, we can iteratively re-examine—at increasing levels of details—the principal chain of events involved in speech inversion. In doing so, we should uncover the terms and concepts that are most necessary.

2.1 Speech sounds as French vowels

We began this report by describing a simple progression from speech sound to vocal tract shape via speech inversion. Let us first address the sounds of speech and their corresponding vocal tract shapes by considering the three cardinal vowels of French, /**i**, **a**, **u**/. Their selection is particularly useful because they entail three articulatory extremes [3]. A phonetic description of these vowels, in the context of all (non-nasal) French vowels, is shown in Table 1.

Table 1: IPA symbols for French vowels

Vowel	Description	Example	
/**a**/	open front unrounded vowel	*dame*	/**dam**/
/**ɛ**/	open-mid front unrounded vowel	*faite*	/**fɛt**/
/**e**/	close-mid front unrounded vowel	*ses*	/**se**/
/**i**/	close front unrounded vowel	*oui*	/**wi**/
/**ɔ**/	open-mid back rounded vowel	*bottes*	/**bɔt**/
/**u**/	close back rounded vowel	*tout*	/**tu**/
/**y**/	close front rounded vowel	*tu*	/**ty**/
/**œ**/	close-mid front rounded vowel	*neuf*	/**nœf**/
/**ø**/	open-mid front rounded vowel	*deux*	/**dø**/

2.2 Relevant speech anatomy

Before describing typical vocal tract shapes, a simple look at vocal tract anatomy will equip us with a helpful vocabulary. Several articulators—the parts of the vocal tract that participate dynamically in speech production—

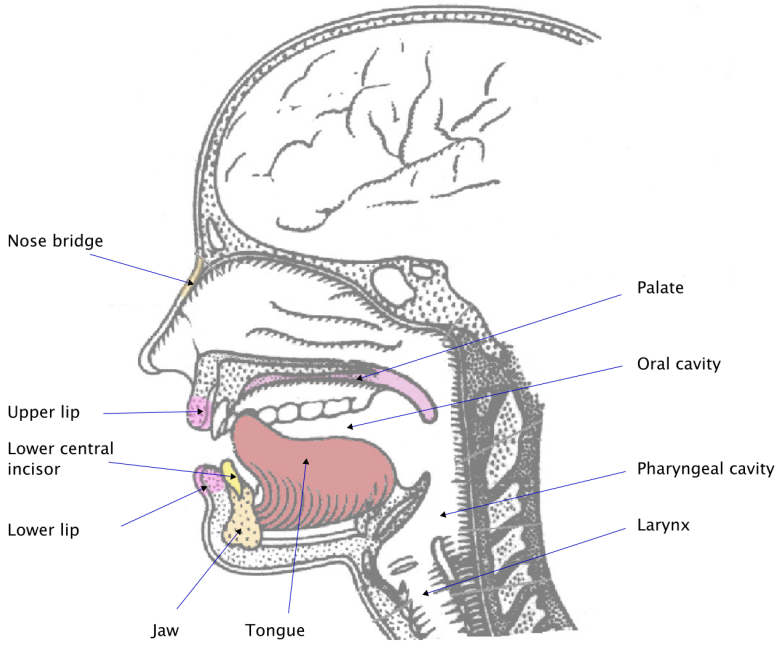

Figure 1: A midsagittal illustration showing many of the anatomical landmarks described in this report.

should already be familiar, such as the lips, tongue and jaw. Relevant regions of the vocal tract include the oral and pharyngeal cavities, along with the larynx, in which the vocal folds are found. Figure 1 illustrates the location of these terms, as well as some anatomical landmarks used later in this report.

2.3 Vowel production

During vowel production, as a speaker exhales, her vocal folds rapidly break the airstream into periodic puffs of air [3]. The complex pattern of this vibration is repeated at a measurable rate called the fundamental frequency, which the human ear perceives as pitch. Harmonic overtones are also pro-

duced, providing additional sound energy at integral multiples of the fundamental frequency. These harmonics are filtered differently for certain sizes and shapes of the vocal tract, resulting in resonance peaks at different frequencies; each resonance peak is called a formant. The position and movement of the speech articulators—such as the lips, jaw, tongue and pharynx—determine the vocal tract geometry and thus influence the distribution of vowel formants. Input to the speech inversion technique at LORIA (for vowels) involves the extraction of the first three formant frequencies, which are sufficient for distinctly characterizing each vowel [12].

2.4 French vowels as vocal tract shapes

We now have a useful base for describing the relationship between vocal tract configurations and the three cardinal vowels /**i**, **a**, **u**/. (Author's note: a significant portion of the three descriptions below draw their information from [3], with input also from [12].)

2.4.1 Vowel /i/

During pronunciation of the first vowel, /**i**/, the oral cavity resonates at relatively high frequencies corresponding to the second and third formants (F_2 and F_3). To achieve this acoustic effect, the size of the oral cavity must small. To create this small volume of air, a speaker can fill most of the oral cavity with the tongue. At the same time, the pharynx grows larger since the rear part of the tongue has moved up and forward, away from the pharyngeal cavity. Lips are not protruded during /**i**/; indeed, they are often spread.

2.4.2 Vowel /a/

The vowel /**a**/ provides an opportunity to discuss a phenomenon that is important in this work—the compensatory effect of multiple articulators. Here, two different strategies can generate the larger oral cavity and smaller pharyngeal cavity that contribute to the acoustics of /**a**/. To reduce the size of the oral cavity, a speaker can either (1) passively lower the jaw to

lower the tongue, or (2) actively lower the tongue by contracting an extrinsic tongue muscle. A combination of both strategies is also possible. The resulting configuration is a smaller pharyngeal cavity that resonates to higher frequency harmonics (high F_1) and an enlarged oral cavity that influences a relatively low F_2.

2.4.3 Vowel /u/

To create the vocal tract shape necessary for /**u**/, a speaker typically raises the dorsum (rear) of the tongue toward the palate to a position much further back than the one described for /**i**/. Like /**i**/, the effect of raising is an increased pharyngeal cavity and therefore decreased F_1. At the same time—and quite the opposite of /**i**/—protrusion and rounding of the lips lengthen the oral cavity, resulting in a lower F_2. Again, compensatory effects are possible: by lowering the larynx, the speaker can achieve the same effect as lip rounding and protrusion.

Figure 2 shows a set of midsagittal MRI scans depicting typical configurations for the pair of vowels /**i**, **u**/.

2.5 Articulatory measurements

2.5.1 Imaging techniques

These descriptions of vocal tract configurations are the result of years of data-driven observations and analyses in the speech research community. Older data collection efforts tended to acquire midsagittal X-ray images, which provide full-length tongue imaging (and possibly the whole vocal tract) at relatively slow time resolutions. Safety concerns, however, regarding exposure to ionizing radiation mean that old collections of images are still in contemporary use [18][14], due to the absence of recent collection efforts.

An alternative imaging technique includes magnetic resonance imaging (MRI), which provides excellent spatial resolution and full-length imaging of the vocal tract, but insufficient temporal resolution. Images are also taken

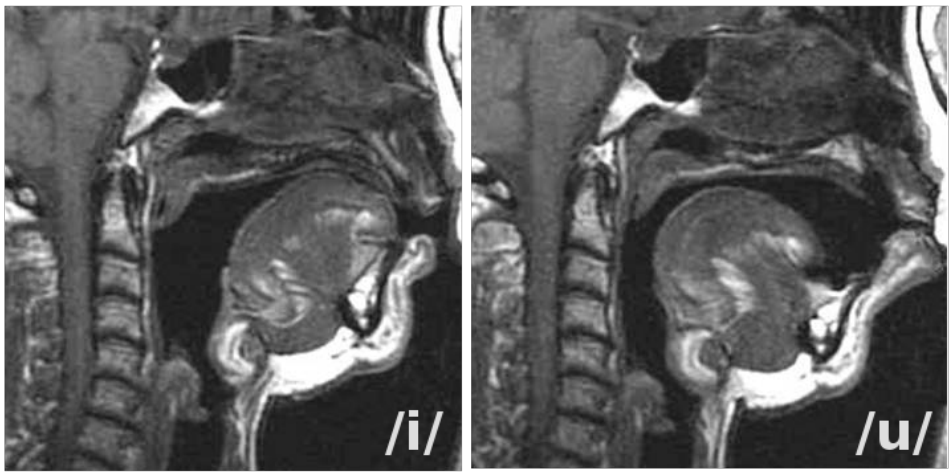

Figure 2: Midsagittal MRI images of the two indicated vowels. Note the correspondence between the cavity areas in the images and their description in the text.

along the midsagittal plane (2D), although the construction of static 3D volumetric images is possible.

Ultrasound imaging is inexpensive and portable. It can provide reasonable time resolution, as well as full-length visualization of the tongue, but at decreased spatial resolution. The lack of an absolute spatial reference introduces a significant challenge, although research is underway at LORIA to use this technique to evaluate speech inversion.

A fourth approach to measuring the shape and position of speech articulators is electromagnetic articulography (EMA), described in the next section. In this present study, EMA forms the basis for collecting natural vocal tract configurations over time. It offers excellent time resolution and the latest articulograph models sample data over a 3D measurement space.

2.5.2 Electromagnetic articulography

Speech research and its applications often face the challenge of measuring and visualizing movement within the vocal tract. Through a series of fixed transmitters and carefully placed sensors, Electromagnetic Articulography

(EMA) allows the concurrent measurement of external articulators, such as the lips and lower jaw, along with certain internal articulators, such as the tongue tip, tongue body and—with more difficulty[1]—the velum. Movement of the tongue root, the pharyngeal muscles and the larynx, meanwhile, remain beyond the reasonable placement of EMA sensors. As such, EMA presents a useful tool for measuring lip, jaw and tongue position for French vowels.

The most current articulograph model, the Carstens AG500, involves 12 electromagnetic receiver coils (sensors) and six transmitter coils. The transmitter coils are fixed around a speaker's head, creating a spherical measurement area that is 300 mm in radius. Within this area, the speaker can move freely, since the position of two or more head-placed reference sensors allows later correction for head movement. The remaining sensors (for a total of 12) can be placed on accessible speech articulators, such as the lips, lower jaw and tongue. Figure 3 shows a subject sitting within the measurement area of the AG500.

During operation of the articulograph, the six transmitter coils generate alternating electromagnetic fields, each with a specific frequency. In response, an alternating current is induced in each of the 12 sensors. The strength of the induced current is inversely proportional to (approximately) the cubed distance between the transmitter and the receiving sensor [7]. From these measured current strengths—and since each transmitter emitted a different frequency—software accompanying the AG500 can compute the three-dimensional coordinates of each sensor's position within the measurement area, as well two angles of orientation (its tilt and yaw). The sampling rate of these measurements is 200 Hz, with spatial resolution claimed to be better than 1 mm [20].

An earlier EMA system, the Carstens AG100, while able to sample data at 500 Hz, forced measurement along a midsagittal plane and therefore could not track lateral movement of the tongue, nor determine lateral dimensions

[1]Hoole and Nguyen [8] suggest that sutures may be required to ensure reasonable fixation of the electromagnetic sensors to the velum. On the other hand, Richmond [17] includes the velum as a sensor position in his data.

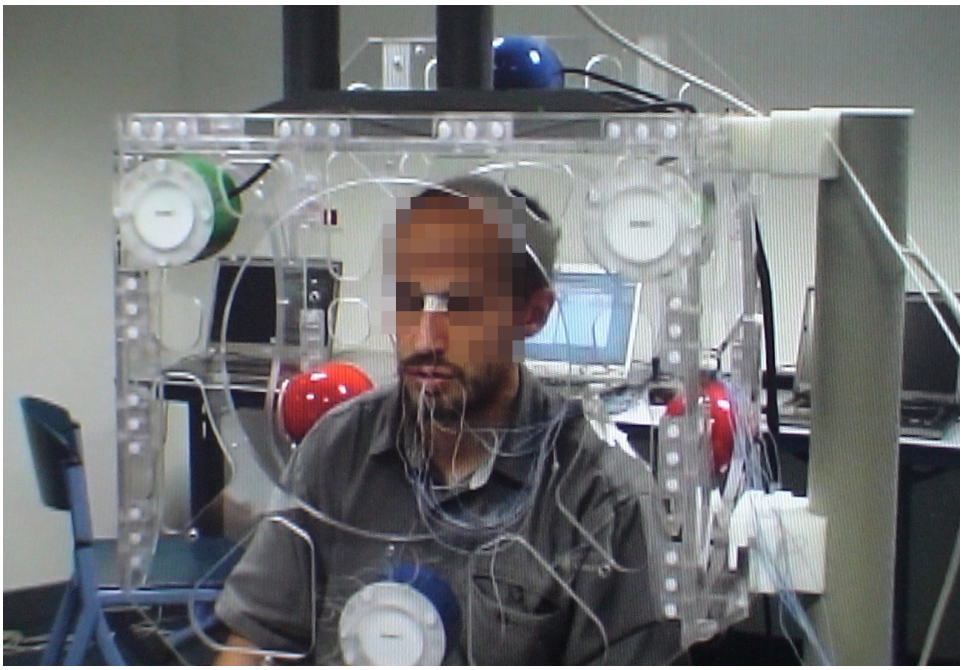

Figure 3: A speaker sitting within the measurement area of the Carstens Articulograph AG500. A transmitter coil is encased within each of the six coloured balls surrounding the subject (three to his front and three to his rear). The receiver coils (sensors) are placed on different fleshpoints—in this photo, most visibly above the nose bridge and on the mouth, but also on the tongue. Each fleshpoint sensor is connected to the AG500 by a thin, flexible filament.

of the mouth opening. To facilitate this setup, the speaker's head was necessarily constrained within a helmet [21]. In contrast, the free movement permitted by the AG500 provides improved speaker comfort and encourages greater naturalness in her speech.

In a sense, EMA shares a similar goal with speech inversion: the visualization of articulator movement during speech. Although the primary aim of EMA involves measurement, recent research in speech pathology has investigated the benefit of using EMA for visual feedback during remedial therapy [11], [10]. To this end, a fully functioning speech inversion system could serve as an inexpensive[2], less invasive and less time-consuming approach to recovering vocal tract configurations from a speaker. Meanwhile, as speech inversion continues to develop toward its ultimate aim, EMA should offer a way to measure the accuracy of inversion mappings—particularly with respect to tongue position and shape. Establishing a procedure for such an evaluation is the main purpose of this thesis work.

At this level of understanding, we can appreciate the central role of EMA data in the current investigation. More complete technical descriptions of the AG100 and AG500 can be found in [1], [20] and [21].

2.6 Maeda's articulatory model

To generate its codebook mappings, the inversion approach at LORIA uses an articulatory speech synthesizer (cf. [15]). This synthesizer is based on a 2D articulatory model, developed by Maeda [13]. An approximate transformation reconstructs the vocal tract in 3D space [2] (cited in [16]) to provide a more realistic model of acoustic output.

2.6.1 Articulatory parameters

Maeda's model describes the dynamic form of the vocal tract through a weighted sum of seven articulatory parameters. These parameters were isolated by guided principal component analysis (guided PCA) from hand-

[2]According to the manufacturer's web site, as of December 2007, the catalog price for the AG500 is just under 80,000 euros.

drawn contours on a series of midsagittal X-ray images. The images were taken in the 1970s, during the production of French vowels by an individual female speaker, PB. Each of Maeda's parameters can vary between -3 and $+3$, a range that restricts the vocal tract configuration to most possible shapes and thus avoids unrealistic and impossible forms. The approximate direction of variation of these parameters is illustrated schematically in Figure 4. Together they account for more than 98% of the total variance in speaker PB's vocal tract shape during vowel production.

2.6.2 Maeda's grid

In Maeda's articulatory model, the shape of the vocal tract walls along the midsagittal plane is represented by two contours, plotted against a semipolar coordinate system. The origin $(0, 0)$ of this system is aligned to a fixed point on the lower jaw. Three subgrids, meanwhile, contribute to the overall coordinate system: two Cartesian subgrids overlay the oral cavity and the pharyngeal cavity (which also includes the upper laryngeal region); between them, a polar grid is superimposed on the oropharyngeal region. Figure 5 illustrates these regions.

The spacing between each grid section depends on the region of the vocal tract model and the speaker it represents. For the original female speaker, gridlines in the two Cartesian zones area spaced 0.5 cm apart. In the polar grid region, coordinate gridlines are drawn at 11° intervals about the origin. For new speakers with different vocal tract dimensions, the original model must be appropriately scaled in each of the three regions, as we will see later.

With this composite coordinate system in place, the shape of the vocal tract can be described by a set of two vectors. One vector contains the coordinate values of the intersection points between the upper wall contour, while the other vector contains those for the lower wall contour. Each vector has a length of 31, corresponding to the total number of grid sections. The stored values are the distances along each section that any intersection points lie. In essence, these vectors describe the vocal tract shape, as it is sampled

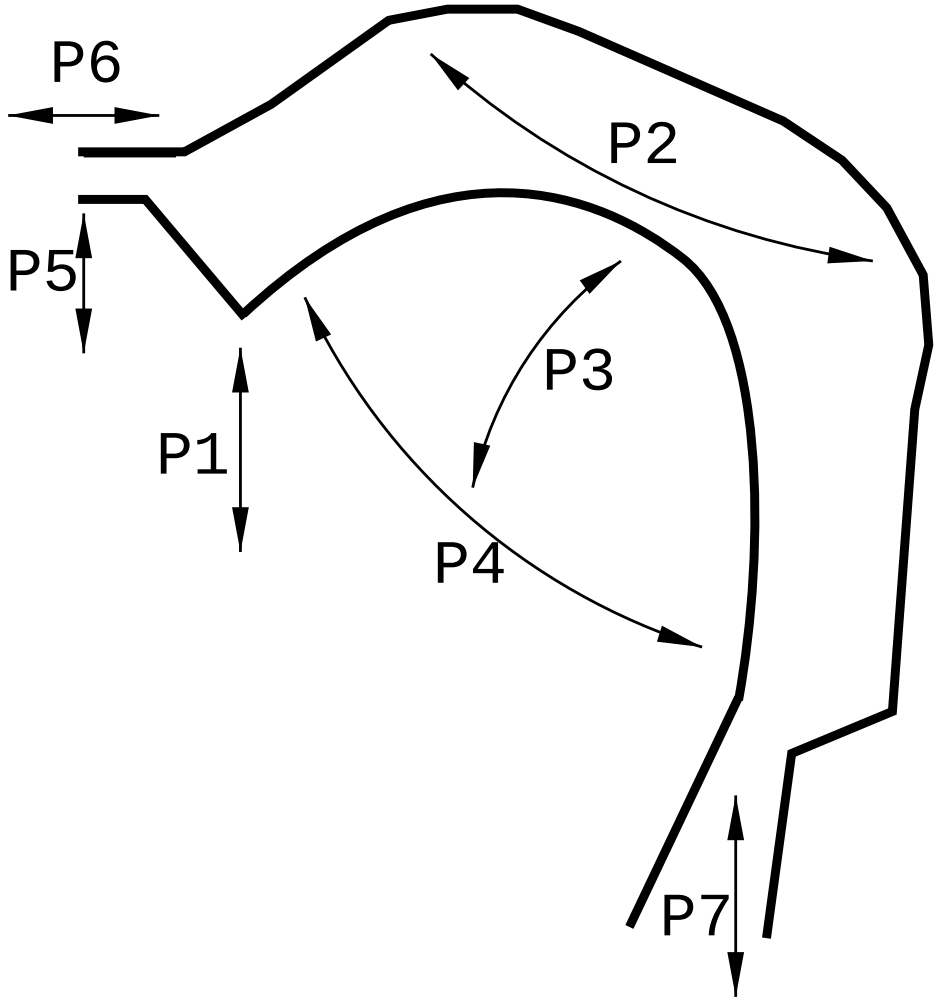

Figure 4: Direction of variation for the seven parameters of Maeda's articulatory model: P_1 (jaw position), P_2 (tongue body position), P_3 (tongue body shape), P_4 (tongue tip position), P_5 (lip opening), P_6 (lip protrusion), P_7 (larynx height).

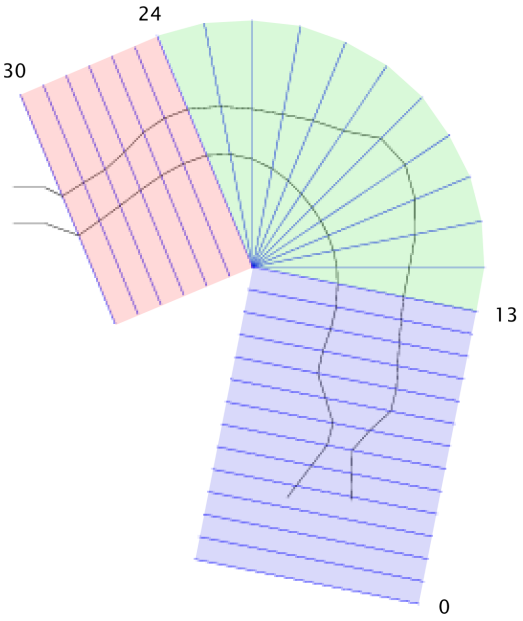

Figure 5: The coordinate system of Maeda's articulatory model. The upper region (red) corresponds to the oral cavity and the lower region (blue) corresponds to the pharyngeal cavity (and includes the larynx). Both use a Cartesian coordinate system. The middle grid (green), which follows a polar coordinate system, is the oropharyngeal region. Grid section numbers are indicated.

at each coordinate gridline.

2.7 Speech inversion using a hypercube codebook

At an abstract level, acoustic-to-articulatory inversion can be described as a technique that receives sound input from a speaker and, as output, predicts the speaker's vocal tract shape. To make such a prediction, the technique refers to a complex look-up table or codebook, built by an existing articulatory speech synthesizer and based on an established 2D articulatory model (Maeda's model).

A challenge to recording these mappings, however, is the nonlinear relationship between articulatory movements and acoustic output: with certain configurations, a very small change in articulator position can result in a very large change in acoustics—and vice versa. The construction of the codebook accounts for this property [15].

For the acoustic input, the inversion process receives a set of acoustic properties (e.g., formant values for vowels) derived from the sound signal and uses the codebook mappings to propose possible sets of corresponding articulatory parameters. A Viterbi-like dynamic programming technique constrains the best possible solution to a realistically smooth trajectory of vocal tract shape over time (Figure 6). This final step is important because, due to the compensatory effects of different speech articulators, different vocal tract configurations can yield the same acoustics.

2.8 Research goal

If future research efforts are to meet the inherent challenges of speech inversion, success will depend on an ability to measure the accuracy of the acoustic-to-articulatory mappings. This strong need for improved evaluation techniques underlies the goal of the work presented here. Specifically, our aim is to contribute to the development of an evaluation technique that draws from a test corpus of natural EMA data, focusing on the shape and position of the tongue. The scope of the acoustic input that we consider for speech inversion is a subset of French vowels, focusing mainly on /**i**, **a**, **u**/

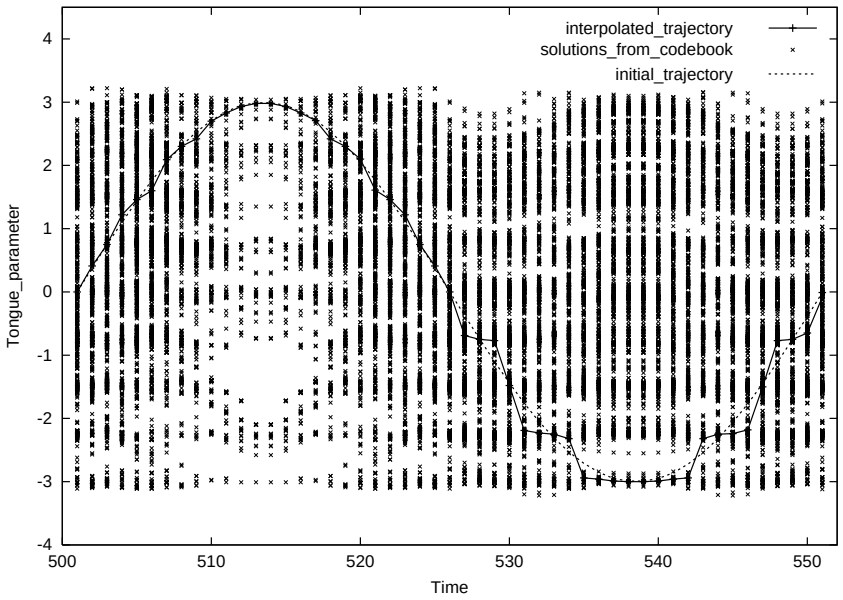

Figure 6: For a given articulatory parameter, a Viterbi-like dynamic programming technique constrains the best possible solution to a realistically smooth trajectory of vocal tract shape over time. (Plot taken from [15].)

(even though data were collected for all non-nasal vowels in French).

An important stage in the direction of this work, but beyond its current scope, is the direct comparison of inversion results to a corpus of real-world vocal tract shapes registered over time—using the same speaker and the same acoustic signal. To facilitate such an evaluation, we need to represent these EMA data in the same currency of output used by the existing speech inversion technique at LORIA. This output, a proposed vocal tract configuration, is expressed in terms of articulatory parameters, based on Maeda's 2D articulatory model. Therefore, the overarching goal of this project is to convert a set of EMA sensor positions in 3D space to a corresponding set of these model parameters.

3 Method

3.1 Overview

The evaluation-by-EMA process begins with the collection of articulograph data and the simultaneous recording of the speech signal. Software accompanying the Articulograph AG500 uses head reference points to correct for overall head movement during speech. Next, extraction of the relevant 3D coordinates occurs for each sensor at each time frame. Using two different approaches, we calculate the midsagittal plane, which allows projection of the 3D data onto a 2D coordinate space. Transformation of the data into 2D space is necessary since Maeda's articulatory model—on which the speech inversion codebook at LORIA is built—assumes a 2D vocal tract. Next, we adapt the model to the vocal tract dimensions of our speaker. Once the real-world data is aligned with the model, we interpolate a cubic spline through the tongue data points. The intersection points of this curve with the coordinate grid of the articulatory model allow us to determine the model parameters that correspond with the original speech articulation. To obtain these parameters, we constrain them within the model's allowable range and solve a system of equations by quadratic programming. The solved model parameters effectively describe the vocal tract configuration associated with the real-world articulograph data at each time frame. In future work, we can evaluate our speech inversion technique by providing it with the same acoustic input from the EMA recording session—and then compare its output parameters with the previously determined parameters; i.e., the parameters that we dervied from the EMA data. By measuring the similarity of these sets of parameters—or indeed the similarity of the tongue shapes they describe—we have a potentially useful measure of how well the speech inversion technique predicts realistic vocal tract shapes.

3.2 Collection of 3D articulograph data

We begin with the collection of articulograph data for the lips, jaw and tongue along with the simultaneous recording of the speech signal. Data

for this investigation benefited from a scheduled collection of EMA data at LORIA for testing and developing algorithms and processing scripts. The simultaneous audio recordings were sampled at a rate of 16 kHz.

Prior to the recording session, a square piece of silk (approximately 5 mm $\times$ 5 mm) was glued to each sensor to increase its surface area. This step aimed to improve the strength of each sensor's attachment to our speaker's fleshpoints and thereby reduce the occurrence of sensor repositioning during the session.

We identify our speaker as YL, an adult male who is a native speaker of French. He had previous experience articulating speech sounds while being measured by an articulograph.

Three of the 12 available sensors were employed as head reference points on YL: one sensor was placed between the eyes (just above the nose bridge), while the other two sensors were each placed behind each ear. The remaining nine sensors were divided among the lips, jaw and tongue. Four sensors outlined the mouth opening: one at each corner, accompanied by sensors on the upper and lower lips. To capture jaw movement, a single sensor was placed on one of the lower central incisors, since gluing directly on a tooth—rather than between two teeth—appeared to offer a stronger attachment. Four sensors were carefully set along the midsagittal line of the tongue, beginning at the rear with the tongue dorsum and followed in the anterior direction by placement on the tongue body, the tongue blade and the tongue tip. In fact, sensor attachment began with the tongue sensors, since these are arguably the most difficult to attach. Figure 7 shows a schematic representation of these sensor locations.

Data processing limitations of the AG500 require that the recording session be divided into several parts, termed *sweeps* [5], which correspond to separately numbered files of collected speech data. This division is in fact convenient for organizing the reading of the corpus, since there was no goal to collect more than two minutes of speech at a time. During the second of 15 sweeps of speech data (i.e., sweep 0002), a series of 36 vowel-vowel (VV) sequences was pronounced by speaker YL over an approximately 60-second period. A mispronounced VV sequence was immediately repeated

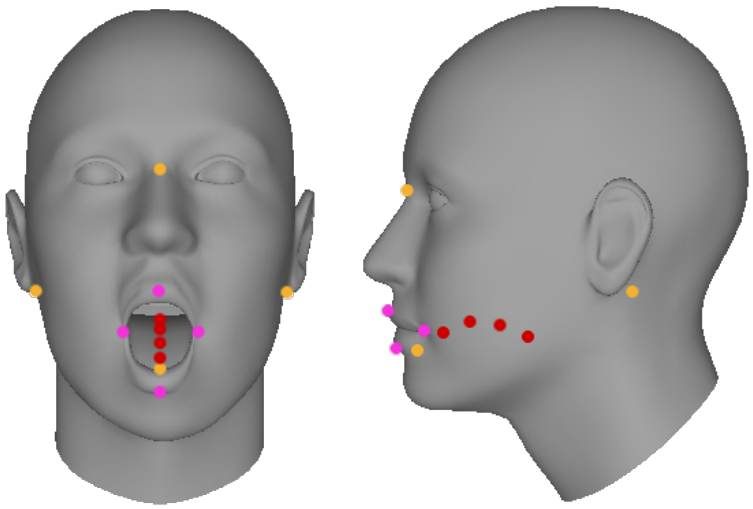

Figure 7: Placement chart of the EMA receiver coils (sensors) on our speaker. Pink dots show the four sensor locations on the mouth. Red dots show the placement of the four tongue sensors. Gold dots indicate the location of the head reference sensors: above the nose bridge, behind each ear and on a lower central incisor.

if the speaker or the experimenter determined this necessity—for instance, when a glottal stop was unintentionally introduced between the two target vowels. Other criteria for repetition included situations where the speaker did not quite produce the intended vowel or did not sustain it long enough. Sweep 0002 contained four repeated corrections.

In our speech corpus, if we denote a VV sequence as V_1V_2, then V_1 included the vowels /**a**, **i**, **y**, **u**/ and V_2 included /**ɛ**, **e**, **i**, **ɔ**, **o**, **u**, **y**, **œ**, **ø**/. Taken from sweep 0002, the 36 combinations of these two vowel sets (e.g., /**aɛ**, **ae**, **ai**, ..., **uy**, **uœ**, **uø**/) form the core set of speech data for developing an evaluation-by-EMA protocol.

Aware that collected EMA data points would later need to be aligned to an articulatory model, a second phase of the recording session involved tracing the midsagittal contour of the palate. For this purpose, a sensor was removed and glued to the end of a pen. Once in place, the speaker slowly moved the pen-sensor along his palate from back to front. This procedure was repeated four times, each in separate sweeps.

During a recording sweep, for each of the 12 sensors, the AG500 records the electromagnetic amplitude associated with each of the six transmitter coils. Software accompanying the articulograph, called CalcPos, uses these six measurements to determine the sensor's position in three-dimensional space, as well as its two orientation angles [5]. (Effectively, CalcPos solves for five unknowns from six equations.) Next, a Carstens software tool named NormPos uses data for the three head reference points to generate a normalization pattern and correct for head movement during speech [4].

A plot of the 3D data collected from our speaker for an instance in time is shown in Figure 8.

3.3 3D to 2D data transformation

Recall that, for each time step, the output of the speech inversion method at LORIA is a set of parameters from Maeda's articulatory model. For the collected EMA data to be useful in the context this 2D model, we therefore need to transform the 3D positions of the lip, incisor and tongue sensors, so

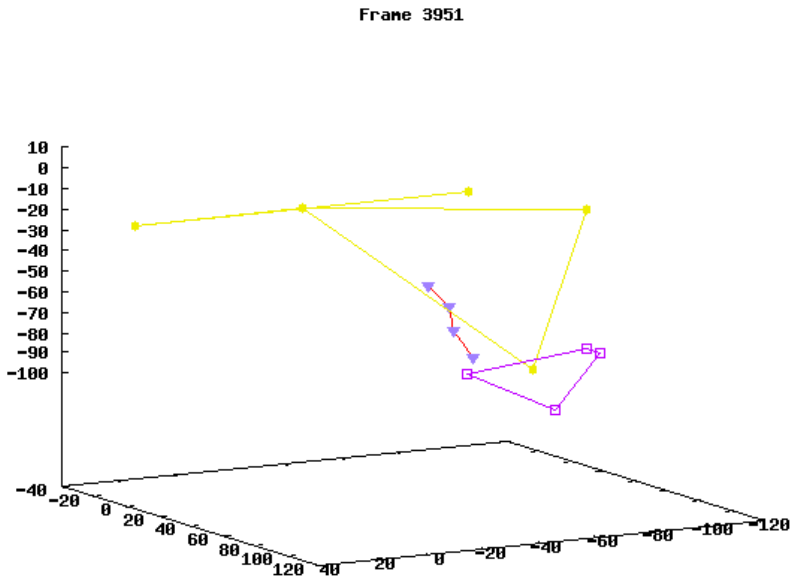

Figure 8: A 3D plot of EMA sensor positions for speaker YL. The purple triangles connected by the red line correspond to the four sensors place on the tongue. The pink diamond outlines the mouth, while the yellow plot illustrates an approximation of the midsagittal plane (as a yellow triangle), calculated by using the midpoint between the sensors placed behind the ears (a straight yellow line). In effect, we are looking at the front of the speaker, but from slightly to his right.

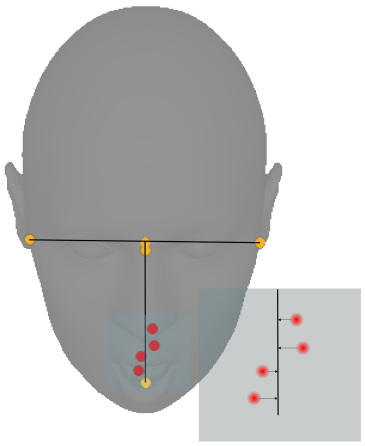

Figure 9: A schematic illustration of the midsagittal projection. The vertical line represents the midsagittal plane, seen from above.

that these nine points lie on the 2D midsagittal plane. By accomplishing this task, we are one step closer to deriving the model parameters corresponding to the measured EMA data.

Figure 9 shows a schematic illustration of projection of the four tongue points onto the midsagittal plane. This data transformation involves the projection of each data point onto a definition of the midsagittal plane in 3D space. An approach to obtaining this definition employs the seven sensors already (approximately) on the midsagittal plane, as we will see in the next section.

3.3.1 Defining the midsagittal plane

The four sensors on the tongue, the sensor on the lower incisor and the two sensors on the upper and lower lips—these seven points already lie approximately on the midsagittal plane of our speaker, with a certain amount of deviation due to their manual placement. For an entire recording sweep (over several thousand samples), we can apply principal component analysis (PCA) to the 3D data points belonging to this set of seven sensors. By calcu-

lating the first two principal components, we obtain a pair of vectors that (in the case of sweep 0002) describes 99.334% of the variance in these sensors' positions. Some 83.696% of the variance is described by the first principle component. Upon visual inspection, this vector appears to coincide with motion along the anterior-posterior axis of our speaker. Meanwhile, the second principle component describes 15.639% of variance, along the speaker's superior-inferior axis. A very small 0.66567% variance in the presumably lateral-medial direction is described by a third and final principle component, which we do not use.

For our recording sweep, we obtain the following two unit vectors from this analysis, where $\mathbf{u}_1$ and $\mathbf{u}_2$ express the first two principal components:

$$\mathbf{u}_1 = (-0.89033, 0.11933, 0.43940) \tag{1}$$

$$\mathbf{u}_2 = (-0.439190, 0.029491, -0.897910) \tag{2}$$

By providing the appropriate 3D EMA data points as input, an existing GNU Octave script from the LORIA Parole team generated these values.

3.3.2 Definition of a new x-axis and y-axis

At every time step, we wish to project the tongue data points onto the midsagittal plane. This action is equivalent to a projection from the original 3D coordinate space (x, y, z) onto a new 2D coordinate space (x', y'). We can define this 2D space by a pair of orthogonal unit vectors. Using the PCA approach, we can directly use the first two principal components $\mathbf{u}_1$ and $\mathbf{u}_2$ as our pair of unit vectors, since they are already normalized.

3.3.3 Projection onto the midsagittal plane

With our two unit vectors defining our speaker's midsagittal plane, we can proceed with the midsagittal projection. By calculating the inner product between any articulator's data point and these two vectors, we effectively project each 3D data point onto the 2D midsagittal plane. If a data point

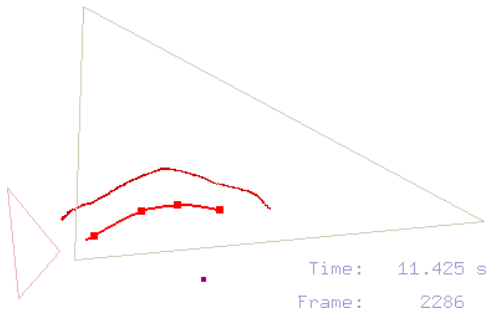

Figure 10: The resulting projection onto the 2D midsagittal plane, as shown in an alignment tool developed in C/C++ for the purpose of this thesis work. The four EMA data points corresponding to the four tongue sensors are indicated as filled squares. For reference, the smaller triangle corresponds to mouth area (upper lip, lower lip and a lip corner), while the larger triangle defines the midsagittal plane via the sensors placed on the lower incisor, the nose bridge and two sensors placed behind each ear (shown here as the midpoint between these two points).

is originally $P = (P_x, P_y, P_z)$, its projection on the x-axis and y-axis of the midsagittal plane is, respectively,

$$P_{x'} = P \cdot \mathbf{u}_1 = P_x u_{1x} + P_y u_{1y} + x P_z u_{1z} \tag{3}$$

$$P_{y'} = P \cdot \mathbf{u}_2 = P_y u_{2x} + P_y u_{2y} + x P_z u_{2z} \tag{4}$$

where $(P_{x'}, P_{y'})$ is the newly projected data point, expressed in a 2D coordinate system.

To summarize, the result of this projection, when applied to the lip, incisor and tongue points, is a description of these points in 2D coordinate space, in the form of a slice along the midsagittal plane (see Figure 10). As such, the transformed EMA data points lie in the same coordinate space as Maeda's 2D articulatory model. At this point, however, their relative position is not yet aligned with the model grid, nor does the model match the dimensions of our speaker YL.

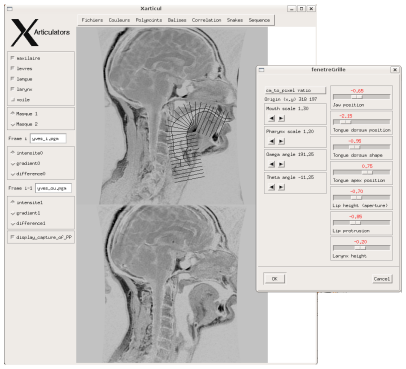

Figure 11: A screenshot of an existing software tool, Xarticul, developed at LORIA—being used here to determine the scale factors for speaker YL.

3.4 Speaker adaptation of the model

Since Maeda's articulatory model was based on data from an individual female speaker, we need to adapt the model's geometric dimensions to our own male speaker, YL. We achieve this match using a LORIA Speech software tool, Xarticul, which superimposes the model grid onto a midsagittal MRI image of YL (see Figure 11). Through a series of manual adjustments—facilitated by sliders and other GUI controls—we visually match the shape of the upper and lower walls of the model vocal tract to those of the MRI image. For speaker YL, we achieve a good match with a mouth scale factor of 1.30 and a pharynx scale factor of 1.20, with respect to the model's original dimensions, for the cardinal vowels /**i**, **a**, **u**/.

3.5 Alignment to the model

The origin in the EMA coordinate system lies near the centre of the spherical measurement area and does not necessarily match the origin of the model grid. A subsequent step toward evaluation-by-EMA, therefore, involves aligning the 2D data points of each sensor to the model's coordinate system, so that corresponding anatomical features are aligned (e.g, EMA mouth opening to the model's mouth opening). Before we explore two

different methods of data alignment to this grid, we need to specify the orientation of the model grid itself.

3.5.1 Orienting the model grid

Recall from the introduction that Maeda's grid features three parts. The 2D space in the oral cavity and the pharyngeal cavity is modelled with two separate Cartesian grids, whereas the oropharyngeal region in between is modelled using a polar coordinate system. Since our EMA data collection did not track any fleshpoints in the pharyngeal cavity, we can disregard this region.

Prior to aligning the EMA data, we draw the model grid with its origin at $(0, 0)$. To avoid later data transformations in the oral cavity subgrid, we allow its (horizontal) x-axis to coincide with the x-axis of our display. Subsequent data plots and intersection points in this region can therefore use its coordinate system directly.

The oropharyngeal region is a different matter, since its subgrid is polar. When operating in this area, we need to perform a transformation from Cartesian (x, y) coordinates to polar (r, ϕ) coordinates, where

$$r = \sqrt{x^2 + y^2}, \phi = arctan(\frac{y}{x}) \tag{5}$$

3.5.2 Palate trace alignment

For this approach, a software tool was developed in C/C++ using the OpenGL Utility Toolkit (GLUT). For our particular recording session, the 2D projected data points from EMA needed to be reflected in the y-axis (superior-anterior direction) so that the EMA data faced left—in the same direction as the articulatory model. Once reflected, the first anterior point in the EMA palate trace was automatically aligned with the first point in the model palate contour. Subtle manual adjustments via a keyboard interface permitted rotation of the palate trace until a reasonable match was attained. Additional translations in the directions of the x-axis and y-axis

were also allowed. A manually fitted palate is shown in Figure 12

3.5.3 Alignment of the tongue scatter plot

A second approach to alignment avoids using the palate trace altogether. Asterios Toutios developed a GNU Octave script to display the vocal tract walls of the articulatory model. On this display, a scatter of data points are plotted for all four tongue sensors over the entire recording sweep. Since the early part of this sweep contained a series of vowel-consonant-vowel sequences, our recorded data contains instances of the tongue touching the palate at both alveolar (anterior) and velar (posterior) locations. By assuming that the tongue sensors do not pass through the palate, we can use the extreme range of tongue sensor positions to visually align the EMA data points within the model vocal tract walls. An interactive interface allows the researcher, then, to repeatedly propose a translation and a rotation until satisfactory alignment is reached (see Figure 13).

In both approaches, once alignment appeared to offer a sufficient match, the rotation and translation involved were saved and subsequently applied to all 2D EMA data points for all time frames. In this manner, all data points are in alignment with the model's coordinate system.

Due to an unresolved problem in the C/C++ code for the palate trace method—related to model scaling and to implementation of quadratic programming in a later step—the tongue scatter plot method is assumed for all work described after this data alignment phase.

3.6 Finding the intersection points

Since the positions of the EMA tongue points do not necessarily fall on a grid section of the articulatory model, we must interpolate between these points and instead consider the intersection between the grid and the interpolating curve. With the projected and aligned tongue points as input, a cubic spline function from the publicly available GNU Scientific Library (GSL) allows us to generate such a curve.

Now, because we oriented the oral cavity subgrid with the Cartesian

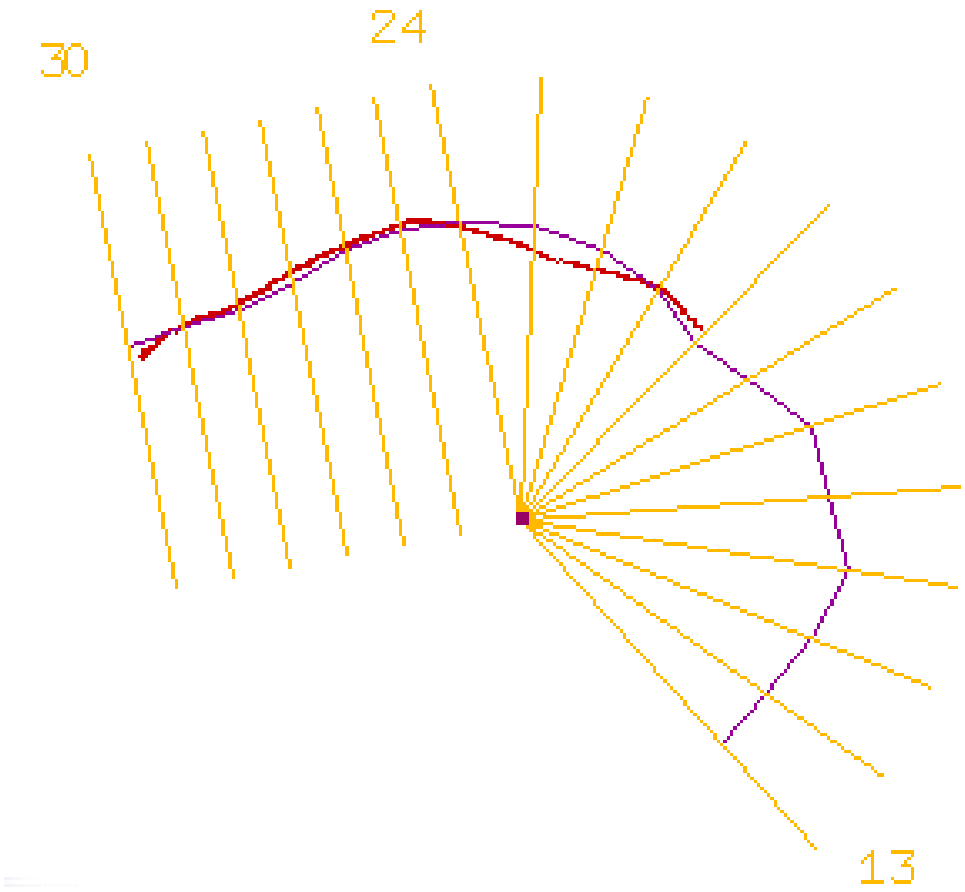

Figure 12: A screenshot from a software tool developed for palate trace alignment. The red contour is the palate tracing and the purple contour is the model's palate. The mismatch in the velar region (soft palate, to the right/rear) is not surprising, since this tissue can raise and lower, whereas in Maeda's model the velum is fixed to a single position.

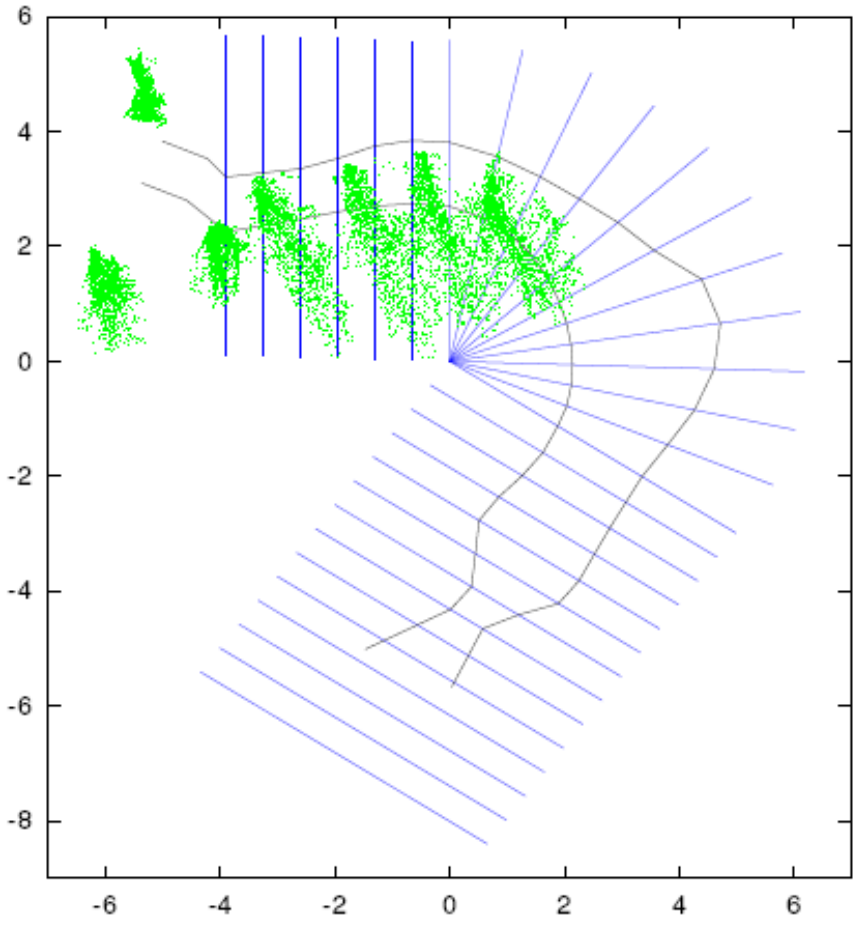

Figure 13: Satisfactory alignment of the EMA tongue data from an entire recording sweep that included extreme movements of tongue. The four green clouds from the right correspond to the range of movement observed for the four tongue sensors. The remaining three clouds to the left are for the lips and lower incisor. Note that we do not allow any of the extreme tongue data points to penetrate the palate.

coordinate system of the visual display, we can obtain the x-coordinate of each intersection point in this region directly from the x-coordinate of each grid section. To obtain the corresponding y-coordinate, we re-evaluate the cubic spline for each of these values of x. For each section of the oral cavity subgrid, we consider this y-value to be the value of interest.

In the oropharyngeal subgrid, which is not Cartesian but polar, we must convert the Cartesian display coordinates (x, y) to a set of polar coordinates (r, ϕ). The determination of intersection points in this space is less straightforward. To do so, we estimate them by least squares, beginning with the Cartesian coordinates of an arbitrary point on the tongue curve and ending with an estimated pair of polar coordinates.

Algorithm 1 Least squares algorithm customized to find the value r that intersects the polar grid at angle ϕ.

$S \leftarrow$ a cubic spline running through the four EMA tongue points
$\phi \leftarrow$ angle for the current section of the polar subgrid
THRESHOLD $\leftarrow 0.001$
MAX_ITERATIONS $\leftarrow 1000$
$i \leftarrow 0$
$d \leftarrow 100.0$
$x \leftarrow 0$
while $d >$ THRESHOLD and $i <$ MAX_ITERATIONS **do**
    $y \leftarrow evaluate\_spline(S, x)$
    $x' \leftarrow \frac{y}{tan(\phi)}$
    $d \leftarrow (x' - x)^2$
    $x \leftarrow x'$
    $i \leftarrow i + 1$
end while
$r \leftarrow \frac{x'}{cos(\phi)}$

As outlined in Algorithm 1, having already calculated a cubic spline for the tongue points, we begin by setting a threshold to govern how close we estimate our value for the polar radius r, given angle ϕ for the current polar grid section. At this time, we also set the maximum number of iterations we wish to run for each estimate. We arbitrarily set the initial square difference to 100.0, a value greater than the threshold, and we choose 0 as an arbitrary

initial value for x.

Next, we enter a loop, respecting the threshold and the iteration limit. For each iteration, once we have calculated the y-value of the spline for the current value of x, we then calculate a new value of x that lies on the current section of the polar subgrid. Since our goal is to minimize the square difference between these old and new values of x, we calculate this difference before we both update our current x-value and increment the counter in preparation for the next iteration.

Upon meeting the criteria for exiting the loop, we have a point x' that approximates an intersection between the tongue curve and the current section of the polar grid. We can then transform this value to the corresponding radius, r, since this region of the model operates in a polar coordinate system.

We perform the above calculations for all model grid sections that are intersected by the tongue spline. Typically, for a given time step, we collect a series of four to seven intersection values, depending on the shape of the tongue. A case where the EMA data led to seven intersection points is shown in Figure 14. We next use these values to solve for the model parameters that best match the current tongue spline.

3.7 Solving for the model parameters

3.7.1 Recapitulation

Before we step into the final task of solving the model parameters, a brief review of the current situation is worthwhile. We have transformed the positions of four tongue sensors from the 3D measurement space of the AG500 articulograph into 2D space. This new coordinate system is consistent with the dimensions of Maeda's articulatory model—and both occur on the speaker's midsagittal plane. The model's geometry, however, must be adapted to that of our speaker. Our procedure also requires that we align the EMA data to the orientation of the model's grid. Once aligned, we fit an interpolating curve between the four 2D tongue points. To express the location of the tongue with respect to the model, we then calculate any

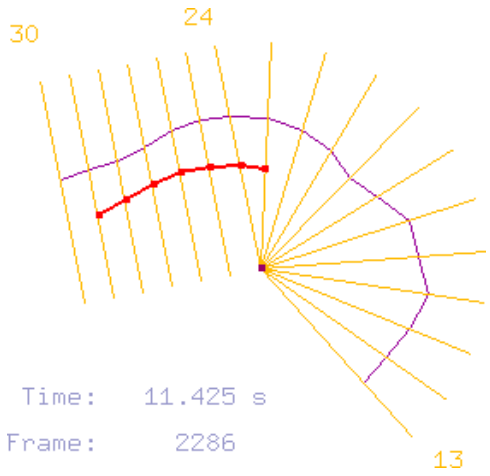

Figure 14: A series of seven tongue intersection points with the model gridlines, with a cubic spline drawn between them. The origin of the model coordinate system is at the centre of the figure, with coordinate sections 13–30 displayed. The speaker is facing left.

intersection points between the curve and each grid section.

3.7.2 Parameter estimation

Recall now that our aim is to generate a set of parameters for a version of Maeda's articulatory model that is adapted to our speaker YL. Ideally, at every 5 ms, we want the configuration of the model vocal tract to match the shape of YL when he was pronouncing each VV sequence. More realistically, with only four data points describing his tongue shape and position, our aim is to find the best model configuration for the three tongue parameters—P_2 (tongue body position), P_3 (tongue body shape) and P_4 (tongue tip position)—as well as the closely linked jaw parameter, P_1. We therefore make no attempt to solve the remaining parameters—P_5 (lip opening), P_6 (lip protrusion) and P_7 (larynx height)—other than to fix these values at 0.000.

For each EMA data sample (i.e., for each time step), we now have a set of values where the tongue intersected (typically) four to seven sections of the

model grid. For the illustration of our approach, let us assume a set of five intersection values expressed as a column vector **b**. Given these points and a matrix of loading factors (weights) A for Maeda's model, we can solve for the four tongue parameters $P_1, \ldots, P_4$, represented by vector **x**. Returning to the model equation introduced in the first chapter, we had

$$\mathbf{b} = A\mathbf{x} \tag{6}$$

which expands to

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_5 \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,7} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,7} \\ \vdots & \vdots & \ddots & \vdots \\ a_{5,1} & a_{5,2} & \cdots & a_{5,7} \end{bmatrix} \begin{bmatrix} P_1 & P_2 & \cdots & P_7 \end{bmatrix} \tag{7}$$

and, with $P_5 = P_6 = P_7 = 0.000$,

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_5 \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} \end{bmatrix} \begin{bmatrix} P_1 & P_2 & P_3 & P_4 \end{bmatrix} \tag{8}$$

Since we know **b** and A and want to solve for **x**, in effect we have a system of five equations with four unknowns. We can determine these four parameter values in **x** using a the quadratic programming function **qp** in GNU Octave's Octave-Forge library[3]. Using such a function, allows us to specify constraints on the parameter values, which, following Maeda's original model, enforce a range between -3.000 and 3.000 for each parameter.

For every 5 ms in the recording sweep, we have obtained a value for the jaw parameter and each of the three tongue parameters. Our next goal is to visualize the effects of these parameters on the model itself.

[3]Freely available at `http://octave.sourceforge.net`.

3.8 Generation of vocal tract shapes

Existing software tools in LORIA Speech accept Maeda's model parameters and scale factors as input, generating a corresponding vocal tract configuration as output. We can, then, generate 2D midsagittal images of the model for each 5-ms sample of EMA data. An appreciation of these images is important, since—due to compensatory effects—different parameter values can result in very similar (if not the same) vocal tract configurations. In other words, a direct examination of parameter trajectories over time is informative about the individual parameter, but potentially disregards a match in the overall shape of the model. We therefore consider both perspectives when assessing the the fit of the model to the EMA tongue curve.

3.9 Summary

We can summarize the approach to fitting an articulatory model with EMA data as follows:

1. First, we collect the EMA data from this speaker.
2. Then we transform the 3D data into 2D data by performing a midsagittal projection.
3. We adapt the articulatory model to our speaker's dimensions. (Note: this step can occur earlier.)
4. We align the 2D data to the coordinates of the model grid.
5. Then we draw a curve through the tongue points, so that we can find the intersection points with the grid.
6. And finally, we solve for the model parameters.

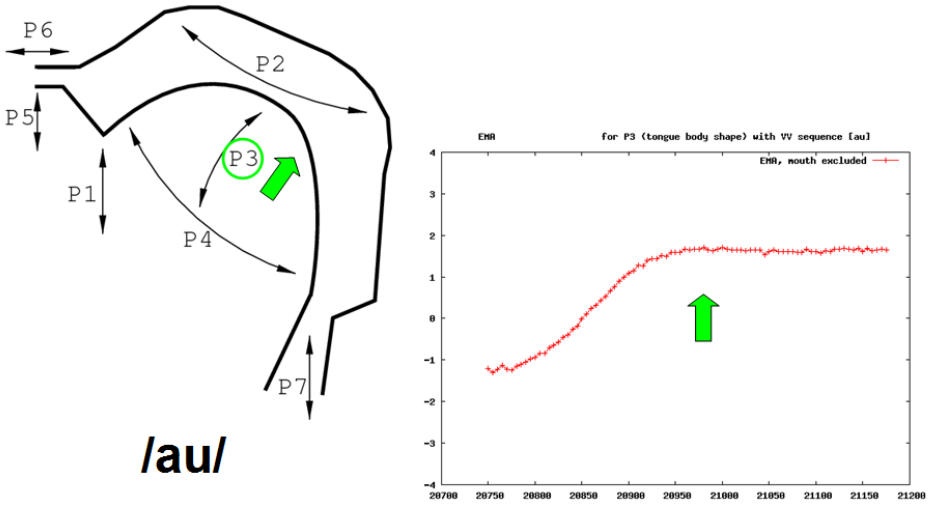

Figure 15: The change in value of P3 (tongue body shape) over time during the vowel-vowel sequence /**au**/. The increase in parameter value corresponds to an increase in tongue height. In the graph to the right, the y-axis is the value of the model parameter (between –3 and +3) and the x-axis is time, expressed in milliseconds.

4 Results

In many instances, a visual inspection of our results reveals a very close match between the interpolated EMA data points on the tongue and the fitted model curve. Model parameters sometimes adopt more extreme values than might be expected, although this result could be explained by necessary compensatory effects due to the fixing of the lip and larynx parameters to default values.

4.1 Parameter results

As described in our procedure, we obtained parameter values over time for different vowel-vowel sequences of French. With a focus on the tongue, we solved directly for model parameters P2 (tongue body position), P3 (tongue body shape) and P4 (tongue tip position). From these values, we inferred

values for P1 (jaw position). We say *infer* for jaw position because the EMA data used to determine parameter values P1 through P4 were tongue data points only. (Recall that the sensor on the lower incisor was used only to correct for head movement and, in one of the alternative approaches for calculating the midsagittal plane.) Values for the remaining model parameters—P5 (lip opening), P6 (lip protrusion) and P7 (larynx height)—were fixed at 0.0.

For example, a result for Maeda's third parameter, which describes the shape of the tongue body, is depicted in Figure 15. For the vowel-vowel sequence /**au**/, we see this parameter increasing during the transition from /**a**/ to /**u**/. This observation corresponds to our phonetic knowledge: for /**au**/ we expect the tongue to move from a low position to a high, back position [12].

4.2 Vocal tract shapes

Examining the trajectories of individual parameters isn't enough, however, to qualitatively evaluate our fitting technique. To better see how well the solved parameters correspond with the original articulatory data, we need to examine the overall vocal tract shapes that correspond with these parameters values. In this vein, Figure 16 shows a vocal tract shape for the sequence /**ia**/. Here the black contours correspond to part of the articulatory model, whereas the red curve is the spline between the EMA data points. In this example, we see a very good fit, one that is typical of our other results. (Beyond the tongue, however, the larynx configuration appears somewhat unnatural.) Note the small oral cavity and the relatively large pharyngeal cavity, as we expect given background knowledge in phonetics. That the general dimensions of the model cavities match our expectations is encouraging, given that we have only solved for four of the seven parameters.

Following the previous image, Figure 17 shows an instance of our speaker pronouncing /**a**/ in the second half of a /**ia**/ sequence. We again see a reasonable fit between the the EMA data for the tongue and model curve. Here, the oral cavity is enlarged and the pharyngeal cavity is restricted, as we expect. We also observe a lower jaw position.

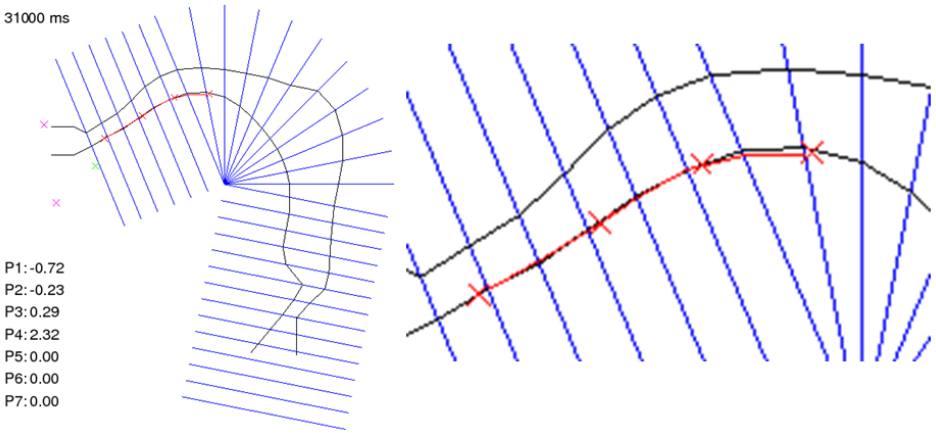

Figure 16: An instance of our speaker pronouncing /**i**/ in the first half of a /**ia**/ sequence. The black contours represent the model, whereas the red contours represent the EMA data points and curve. The overall vocal tract shape is an emergent result of the model, described largely by the seven parameter values listed to the left. P1 is value for the inferred jaw parameter, while P2 through P4 are the three tongue parameters. The parameters for the lips (P5, P6) and larynx height (P7) are fixed at zero. In Maeda's model, all parameter values are constained between –3.0 and +3.0.

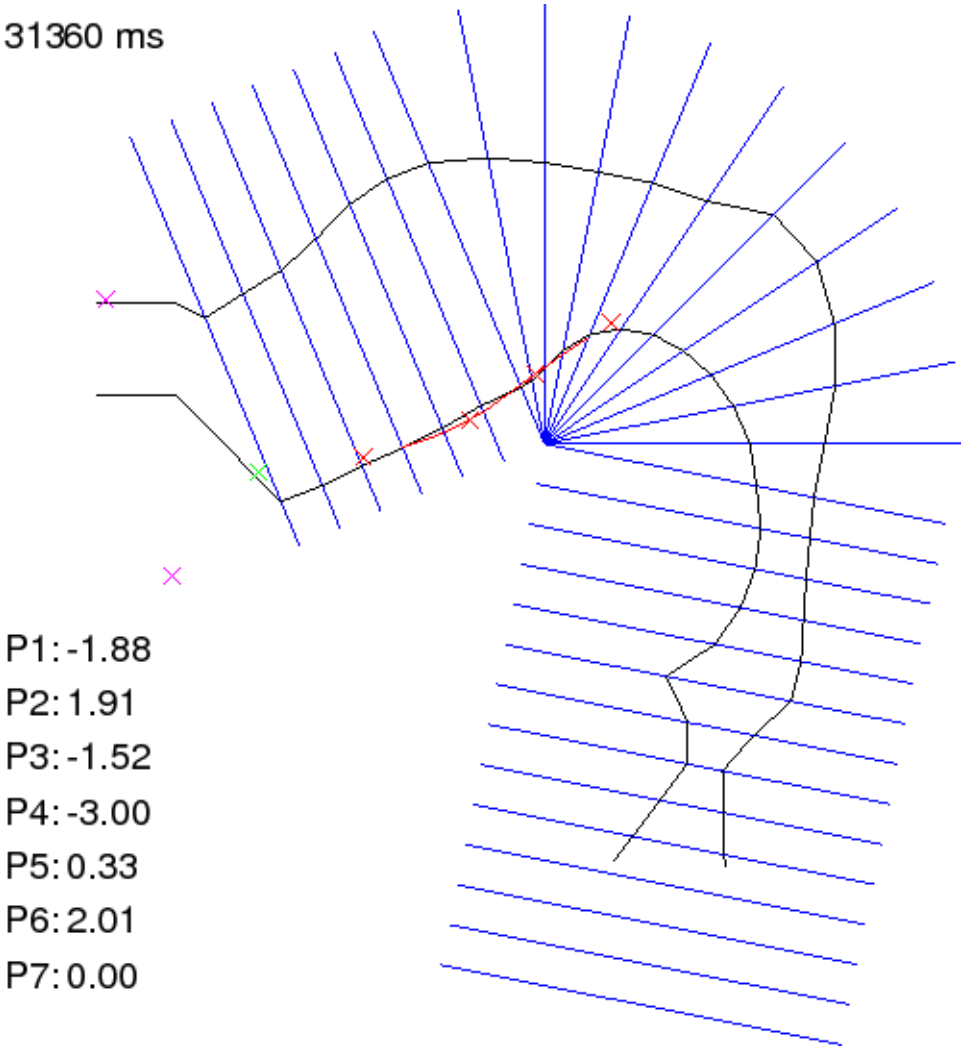

Figure 17: An instance of our speaker pronouncing /**a**/ in the second half of a /**ia**/ sequence.

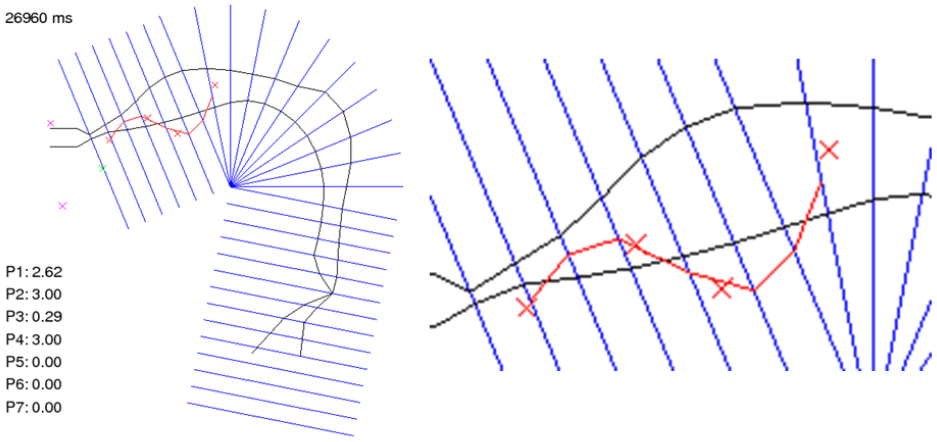

Figure 18: An instance of our speaker pronouncing /**i**/ in the first half of a /**ie**/ sequence. Inspection of the original EMA data revealed that the unrealistic shape of the red contour was due to an intermittent error during the estimation of the sensors' position within the electromagnetic field. Note, however, that the model itself (the black contour) assumes a more realistic tongue shape.

The following observation is not representative of our general results, but we include it as an interesting finding. On one occasion, we observe a very poor fit (Figure 18) due to an error in the EMA data recording, as evidenced in the root mean square value for the original EMA data file (calculated by software accompanying the articulograph). A positive aspect of this particular example is that the model of the tongue appears to be reasonably robust: in spite of the erroneous EMA data, the model contour does not adopt its highly unrealistic position.

4.3 Summary of results

In summary, the observed parameter trajectories tend to match our gross expectations from background knowledge in phonetics. We do observe some

exaggerated articulator positions, especially for the inferred jaw parameter but also in the region of the larynx. Meanwhile, a visual assessment of vocal tract model shapes reveals a good fit between the EMA tongue data and the articulatory model's tongue contour. Similar to the exaggerated parameter values, we find that regions of the model other than the tongue often appear unrealistic. Finally, a positive finding is that the articulatory model appears to be robust to transient noise in the EMA data.

5 Discussion

5.1 Potential sources of error

Implementation of the approach to evaluation presented here revealed several potential sources of error.

- We observed occasional articulograph error when calculating sensor positions. However, such error is detectable through the root mean square error values provided by the articulograph software, which allowed us to avoid these data.

- The impact of the electromagnetic sensors and the presence of wires has a potential impact on the naturalness of a speaker's pronunciation. Our speaker seemed to adapt over time, with obvious effects presence only at the very beginning of the recording session. Research into the use of EMA as visual feedback to remedial pronunciation training suggests that only some speakers are affected in this way [9].

- The definition of the midsagittal plane may be one of our largest source of error, although as a data-driven approach, using PCA seems elegant and robust. We identified an alternative approach using the midpoint between the ears, discussed later.

- Alignment of the EMA data with the model grid is a manual process—whether we use a palate tracing or the tongue scatter plot approach—and is likely another large source of error.

- Fixed parameters for the lips and larynx allowed us to focus on the tongue (as a preliminary investigation), but enhance the compensatory effect on these parameters. The shape and position of the model's tongue, therefore, is effectively encouraged to assume extreme values to achieve cavity shapes and sizes that might otherwise be avoided by protruding the lips or raising the larynx. As we will see in the section on future work, however, incorporating lip data is a next step, as is experimentation with the settings of the larynx parameter.

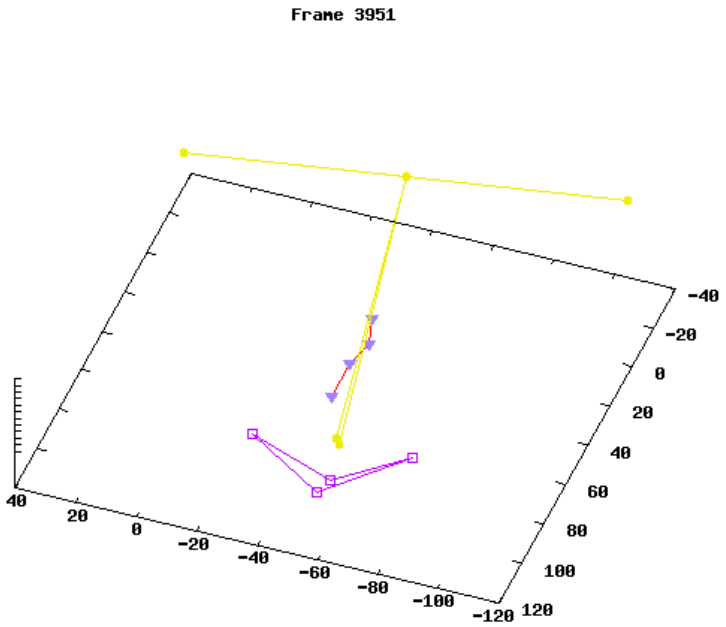

Figure 19: A top view of the same configuration, illustrating the need for projecting the tongue points onto the midsagittal plane.

- A 2D articulatory model assumes that any meaningful speech articulations occur along the midsagittal plane. While this assumption may hold for vowels (our focus), the model may present limitations to other sound classes—especially laterals. Note, however, that this constraint exist only in the model; the use of 3D EMA would remain appropriate for collecting speech movements away from the midsagittal plane.

Finally, in addition to the goal of accommodating a 2D model, Figure 19 shows further motivation for midsagittal projection: a speaker's vocal tract configuration is not necessarily symmetric, nor is the tongue oriented perfectly along the midsagittal plane.

5.2 Conclusion

A general technique for fitting EMA tongue data to Maeda's articulatory model is in place, although variations of certain steps remain areas of further consideration. With just four EMA sensors, we obtained a fit between model and data that, upon visual inspection, appears suitable for continued research in this area. Examining the inclusion of mouth parameters will be important, since we expect a better fit from the additional data, and because we should take advantage of the 3D EMA data we have collected from the mouth fleshpoints. At this point the approach reviewed here is ready to proceed with preliminary comparisons to the results of speech inversion—but awaits multiple refinements in method and tools before being comfortably suitable for frequent and meaningful evaluation efforts.

5.3 Future research

Many avenues for future efforts continue directly from the work presented here. First and foremost, the current set of results for vowels /**a**, **i**, **u**/ warrant closer examination for both good and bad matches between the EMA data and the fitted model. Analysis should include both a visual examination of the dynamic vocal tract configurations, as well as a look at the individual parameters themselves—being vigilant for any unrealistic jumps in parameter values as they change over time. Extreme parameter values and suspiciously constant values should also be on the watch list.

A report of elementary statistics (mean, variance, maxima, minima) for each parameter would be useful to illuminate general trends and problem areas[4]. An examination of maxima and minima, if we determine the frequency at which they are reached (especially for values at the end of the allowable parameter range), could be particular informative and could draw our attention to certain articulatory contexts. To complement a statistical analysis, we should process EMA data across multiple recordings, so that we have multiple instances of each vowel-vowel sequence.

[4]In fact, these statistics have been compiled but have not been analyzed at the time of writing.

Including of the estimation of mouth parameters (lip opening and lip protrusion) is an ongoing research area and we have preliminary results that demonstrate the benefit of considering EMA mouth sensor data in our attempt to improve our fit with Maeda's model. Observing the effects of different larynx parameters should also be considered, perhaps in concert with lip protrusion, since both factors influence overall vocal tract length.

At certain points of the method workflow, we encountered different options. For instance, we could compare the two different grid alignment techniques (palate trace and tongue scatter plot), since this step potentially contributes a significant source of error. A variation in procedure also exists for defining the midsagittal plane: at one point, we tried using a triad of reference points that included the midpoint between the two ear sensors. Revisiting this approach would allow us to compare its worth versus the chosen PCA-based method to the midsagittal plane definition.

On more than one instance, the fit model appears to robust to clear errors in the calculated EMA sensor positions. This area too deserves more attention.

Above all, we desire a quantitative measure of the difference between the EMA tongue curve and the tongue contour of the fitted model over time. A calculation involving root mean squares or the determination of the geometric area between curves might serve as good measures in this regard. Then, we could examine other French vowels (for which the fitted parameters have already been generated) and different phonetic contexts. Division of each vowel-vowel sequence into stable and transition zones has been done and awaits analysis.

5.4 Closing remarks

An approach to the evaluation of speech inversion—one that relies on real-world articulatory data—will help determine the accuracy of speech inversion techniques and indeed whether speech inversion is possible for all speech sounds. The extent to which speech inversion is possible will, in turn, determine which applications can benefit from acoustic-to-articulatory predic-

tions. For instance, a low bit-rate encoding technique may require that speech inversion be possible for all speech sounds (unless one selectively encodes only the parts of the signal that can encoded). On the other hand, pronunciation training or retraining could still receive significant benefit from a restricted set of sound classes (e.g., vowels or fricatives), provided that the goal is to improve the articulation of sounds within these classes. For animators, partial predictions of lip, jaw or tongue movements could still reduce the amount of manual work necessary to animate speech articulators.

By definition, an *accurate* speech inversion technique would provide a *predictive* model for acoustic-to-articulatory mapping in humans. An interesting follow-up question—one that delves into motor theories of speech perception (cf., [6])—might be the following: to what extent would a working implementation of speech inversion provide a *working* model of the relationship between speech acoustics and its articulatory basis?

6 Acknowledgements

Work in this project required contributions from many sources—in the form of software tools and libraries, learning material for obtaining new background knowledge, and collaborative assistance from researchers at LORIA and Saarland University.

Existing tools, scripts and libraries were available in a Linux environment, including C/C++ program code and GNU Octave scripts developed by LORIA Speech researchers. Freely available software tools, such as Gnuplot, and libraries like GSL CBLAS, OpenGL GLUT and GNU Octave-Forge were, thankfully, at my disposal. Countless web sites and forums frequently offered anonymous technical assistance.

A Carstens Articulograph AG500 in LORIA lay at the heart of EMA data collection. **Shinji Maeda**'s articulatory model formed a second core of this work.

Principle among the researchers in LORIA Speech, **Slim Ouni** and **Yves Laprie** deserve my gratitude for accommodating my background, my interests and my working style, as they accepted me in this project. They exposed me to a genuine learning experience and introduced me to the exciting potential of speech inversion.

Asterios Toutios was an invaluable compass, illustrating many mathematical concepts and illuminating several aspects of electromagnetic articulography. Meanwhile, a generous helping hand was always available from **Farid Feiz**, who patiently assisted with setup and technical explanations. **Jonathan Demange**, a fellow Master's student working on the use of ultrasound data, provided a challenging benchmark against which I could measure my progress.

An international student would be out of place without assistance from abroad. I am appreciative for a pair of lengthy discussions with **William Barry** at Saarland University in Germany, who helped steady my keel and provided me with opportunities to explain work under progress. **Ingmar Steiner** also showed encouragement and drew my attention to useful references and resources.

Many years ago, my parents instilled me with a curiosity for science and exploration. If my work here contributes fruit to an applied evaluation of speech inversion, then considerable thanks go toward this early cultivation. **Nadiya Yampolska** provided selfless support and encouragement throughout the project.

I owe a significant portion of gratitude to the region of Lorraine, who supported me financially during the past academic year. I also extend my appreciation to our speaker YL (and, in a test run, speaker SO) for bravely supergluing 12 electromagnetic sensors to various facial and oral fleshpoints.

References

[1] P.J. Alfonso, R.J. Neely, P.H.H.M. Van Lieshout, W. Hulstijn, and H.F.M. Peters. Calibration, validation, and hardware software modifications to the carstens emma system. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)*, 31:105–121, 1993.

[2] D. Beautemps, P. Badin, and R. Laboissière. Deriving vocal tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data. *Speech Communication*, 16:27–47, 1995.

[3] G.J. Borden, K.S. Harris, and L.J. Raphael. *Speech science primer: physiology, acoustics, and perception of speech.* Lippincott Williams & Wilkins, Philadelphia, fourth edition, 2003.

[4] Carstens Medizinelektronik GmbH. *Program Description 04.05: NormPos.exe (Revision 0)*, November 22, 2004. [http://www.articulograph.de/AG500/ag500_man/NormPos.pdf].

[5] Carstens Medizinelektronik GmbH. *Program Description 04.06: CalcPos (Revision 0)*, November 21, 2004. [http://www.articulograph.de/AG500/ag500_man/CalcPos.pdf].

[6] B. Galantucci, C.A. Fowler, and M.T. Turvey. The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13(3):361–377, 2006.

[7] P. Hoole. Issues in the acquisition, processing, reduction and parameterization of articulographic data. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)*, 34:158–173, 1996.

[8] P. Hoole and N. Nguyen. Electromagnetic articulography in coarticulation research. *Forschungsberichte des Instituts für Phonetik und Sprach-*

liche Kommunikation der Universität München (FIPKM), 35:177–184, 1997.

[9] W.F. Katz, S. Bharadwaj, M. Rush, and M. Stettler. Influences of ema receiver coils on speech production by normal and aphasic/apraxic talkers. *Journal of Speech, Language, and Hearing Research*, 49:645–659, 2006.

[10] W.F. Katz, G.C. Carter, and J.S. Levitt. Treating buccofacial apraxia using augmented kinematic feedback. *Aphasiology*, 12:1230–1247, 2007.

[11] W.F. Katz, J.S. Levitt, and G.C. Carter. Biofeedback treatment of buccofacial apraxia using ema. *Brain and Language*, 87:175–176, 2003.

[12] P. Ladefoged. *A course in phonetics*. Thomson Wadsworth, fifth edition, 2006.

[13] S. Maeda. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle and A. Marschal, editors, *Speech Production and Speech Modelling*, pages 131–149. Kluwer Academic Publishers, 1990.

[14] S. Maeda, M.O. Berger, O. Engwall, Y. Laprie, P. Maragos, B. Potard, and J. Schoentgen. Deliverable d1: Technology inventory and specification of fields investigated. Project no. 2005-021324 aspi audiovisual to articulatory speech inversion, ASPI Consortium, 2006. 124 pages.

[15] S. Ouni. *Modélisation de l'espace articulatoir par un codebook hypercubique pour l'inversion acoustico-articulatoire*. PhD thesis, Université Henri-Poincaré (Nancy 1), 2001.

[16] S. Ouni and Y. Laprie. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *J Acoust Soc Am*, 118(1):444–460, 2005.

[17] K. Richmond. *Estimating articulatory parameters from the acoustic speech signal*. PhD thesis, Centre for Speech Technology Research, University of Edinburgh, 2002.

[18] R. Ridouane. Investigating speech production: a review of some techniques. Unpublished manuscript, 2006.

[19] J. Schroeter and M.M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150, January 1994.

[20] A. Zierdt, P. Hoole, M. Honda, T. Kaburagi, and H.G. Tillmann. Extracting tongues from moving heads. In *Proc 5th Speech Production Seminar*, pages 313–316, 2000.

[21] A. Zierdt, P. Hoole, and H.G. Tillmann. Development of a system for three-dimensional fleshpoint measurement of speech movements. In *Proc XIVth ICPhS*, pages 547–553, 1999.

7 Appendix

List of Figures

List of Tables