INTERSPEECH 2009 BRIGHTON

Deriving Vocal Tract Shapes from Electromagnetic Articulograph Data via Geometric Adaptation and Matching

Ziad Al Bawab[1], *Lorenzo Turicchia*[2], *Richard M. Stern*[1], *and Bhiksha Raj*[1]

[1] Department of Electrical and Computer Engineering and School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. 15213

[2] Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA. 02139

ziada@cs.cmu.edu[1], turic@mit.edu[2], rms@cs.cmu.edu[1], bhiksha@cs.cmu.edu[1]

10.21437/Interspeech.2009-589

ABSTRACT

In this paper, we present our efforts towards deriving vocal tract shapes from ElectroMagnetic Articulograph data (EMA) via geometric adaptation and matching. We describe a novel approach for adapting Maeda's geometric model of the vocal tract to one speaker in the MOCHA database. We show how we can rely solely on the EMA data for adaptation. We present our search technique for the vocal tract shapes that best fit the given EMA data. We then describe our approach of synthesizing speech from these shapes. Results on Mel-cepstral distortion reflect improvement in synthesis over the approach we used before without adaptation.

Index Terms: MOCHA EMA data, Maeda Model, vocal tract adaptation, articulatory model fitting, articulatory synthesis

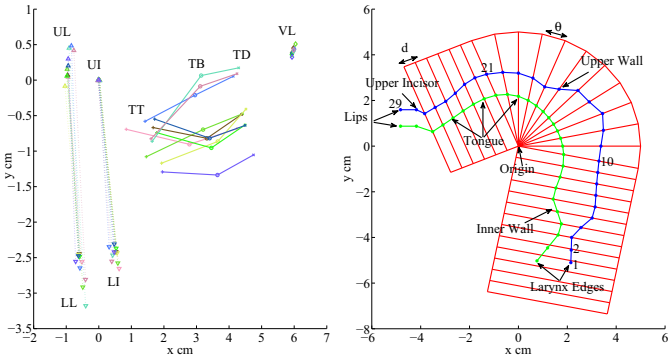
1. INTRODUCTION

ElectroMagnetic Articulography (EMA) has lately been gaining popularity among researchers as a simple technique for measuring the mechanism of speech production [1]. EMA, originally developed in the University of Göttingen in 1982, comprises of a set of sensors placed on the lips, incisors, tongue, and velum of the speaker. A set of transmitters generates magnetic fields at multiple frequencies, each of which induces a current in the sensors. By measuring the levels of generated current, the (x, y) coordinates of each of the sensors can then be determined. Each EMA measurement thus consists of a set of such position coordinates, one from each sensor.

Figure 1.a illustrates the positions of the sensors and the typical measurements obtained from the MOCHA database [1]. As the person speaks, a sequence of EMA measurements is obtained from the sensors. This sequence of measurements is assumed to provide at least partial characterization of the speech production process.

But exactly how reliable are these measurements and how much do they tell us about the vocal tract that produces the speech? The EMA only measures the locations of a very small number of points on the vocal tract, typically four in the location of the lips and incisors, one on the velum, and merely three on the tongue. The vocal tract, on the other hand, is a complex three dimensional object that cannot be fully characterized by a small number of points. Further, the precise location of the EMA sensors themselves is also highly uncertain and impossible to calibrate with respect to the vocal tract. Although the sensors on the tongue are placed at calibrated distances from one another, the elasticity and complexity of tongue structure ensures that their actual positions, both along the tongue surface and relative to overall tongue structure cannot be precisely known.

Given these various factors, it is natural to question the usefulness of these measurements as a characterization of the speech generation process. Clues may be found in work by researchers who have previously shown that the EMA measurements are reliable cues to the speech signal itself. Toda *et al* [2] have produced speech from EMA measurements using learned statistical dependencies between them and the corresponding speech signals, demonstrating that they do indeed relate to the *output* of the speech generation process. Toth and Black [3] experimented with using EMA for voice transformation. While these experiments do provide indirect evidence of the relation of EMA measurements to the speech production mechanism, it is still not clear that they provide direct information about the shape of the speaker's vocal tract itself.



(a) EMA Data (b) Maeda Model

Fig. 1. (a) EMA measurements sampled from the MOCHA database. Notation of EMA data used in this paper is: upper lip (UL), lower lip (LL), upper incisor (UI), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD), and velum (VL). (b) Maeda model composed of grid lines in red, vocal tract upper profile in blue, and vocal tract lower profile in green corresponding to the steady state shape with p_1 to p_7 set to zero.

In this paper we attempt to derive actual characterizations of vocal tract shapes from EMA measurements. Since the EMA itself only comprises a small set of sensor locations, we use a model-based approach to estimate the complete vocal tract configuration from them. Specifically, we use the model proposed by Maeda [4], which represents a mid-sagittal profile of the vocal tract in terms of seven parameters.

One simple approach to arriving at a vocal tract configuration in this manner is to determine the specific set of values for the seven Maeda parameters that best explains the measured EMA sensor positions [5]. This, however, is insufficient. Maeda's vocal tract model is not generic; it was originally developed using 1000 frames of cineradiographic and labiofilm data from only two female speakers. It must be *adapted* to the current speaker. The specific aspects of the

This work was partially supported by NSF (Grant IIS-0420866).

model that are adapted are the height of the palate, the tilt of the oral cavity, and the length of the vocal tract. This is done by comparing the geometry suggested by the ensemble of all EMA measurements for the speaker to that defined by the model. The actual location of individual sensors need not be known; hence the procedure is robust to variations and inconsistencies in sensor placement.

Once the Maeda model is adapted to the speaker, the actual vocal tract configuration corresponding to any set of EMA measurements is obtained through a simple codebook search. We use a codebook of Maeda-model parameters that describes a large sampling of possible vocal tract shapes. For each EMA measurement, we select the vocal tract shape that is geometrically closest to the set of position coordinates represented in it. In order to ensure that the estimate of the vocal tract is based entirely on geometric principles, since the EMA measurements are geometric in nature, we do not use the audio recordings of the speech signal in the adaptation as has been used in previous approaches [6, 7].

The "truthfulness" of the estimated vocal tract configurations can now be evaluated by synthesizing speech from them using an articulatory synthesis model and comparing synthesized speech to the actual speech signal produced during the utterances. We specifically use a modified version of the Sondhi and Schroeter model [8] for this purpose. Experiments show that the synthesized speech is quite similar to the actual speech, both perceptually and in terms of the Mel-cepstral distortion (MCD) metric [2].

2. MAEDA GEOMETRIC MODEL

Maeda's model is composed of a two-dimensional semi-polar grid spanning the midsagittal plane of the vocal tract. The grid is composed of the red lines in Figure 1.b. It is defined by a set of parameters: the *origin*, the width of the grid *d*, and the angle of the polar region of the grid θ. The vocal tract itself is composed of an upper profile and a lower one. The upper profile shown in blue consists of the upper lip and incisor, upper palate, and pharynx and larynx outer wall. The inner profile consists of the lower lip and incisor, tongue, and pharynx and larynx inner wall and is shown in green.

Maeda uses seven parameters to generate the overall profile of the vocal tract. The formulation in Equation 1 summarizes the procedure in pseudo MATLAB code. p_1 is related to the jaw, p_2, p_3, and p_4 to the tongue, p_5 and p_6 to the lips, and p_7 to the larynx. The bases $[B_{larynx}\ B_{uwall}\ B_{tong}\ B_{lips}]$ and offsets $[O_{larynx}\ O_{uwall}\ O_{lips}]$ are derived from the speaker-specific vocal tract profiles Maeda extracted from the 1000 images. These bases are multiplied by the parameters and then added to the offsets to generate different shapes. A 29 dimensional vector is computed using the formulation in Equation 1 and projected onto the grid. The projected points fall on the grid except for the lips and larynx. The vocal tract profiles are composed of the lines joining these points.

Using the seven Maeda parameters with the current model will create vocal tract shapes and generate sounds pertaining to the two speakers from whom the bases and offsets are derived. In order to make the model generate sounds pertaining to different speakers, it has to be able to match their vocal tract shapes. Hence the need to adapt Maeda's model to the EMA data. Since the EMA data are purely geometric, we need to ensure the Maeda's model geometry is accurate enough to be able to characterize the EMA measurements.

Note that the upper palate defined by $UpperWall$ in Equation 1 and shown in blue in Figure 1.b is independent of the seven Maeda parameters and is just a projection of the sum of B_{uwall} basis and the O_{uwall} offset into the geometric grid. Different parameters of the grid lead to different projected shapes. Hence, we choose to adapt the grid parameters and use the bases and offsets without adaptation.

$$
\begin{aligned}
Larynx &= B_{larynx} * [p_1\ p_7]' + O_{larynx} \\
UpperWall &= Proj(B_{uwall} + O_{uwall}) \\
Tongue &= Proj(B_{tong} * [p_1\ p_2\ p_3\ p_4]' + O_{uwall}) \\
Lips &= B_{lips} * [p_1\ p_5\ p_6]' + O_{lips}
\end{aligned} \tag{1}
$$

3. VOCAL TRACT MODEL ADAPTATION

3.1. Origin

We follow an approach similar to the one by McGowan [6] by superimposing the EMA data into the Maeda semi-polar coordinate space. We first need to match the coordinates of the two systems. In MOCHA, the sensor placed on the upper incisor is used as the origin [1]. In Maeda's model, the upper incisor is at a fixed location. Hence we translate Maeda's model coordinates such that the new origin coincides with the upper incisor.

3.2. Upper Wall

Adapting the upper palate ensures that the EMA measurements for the tongue do not extend beyond it. First we estimate the upper wall of the EMA data using the distributions of the sensors positions for all the frames available for the speaker. We compute a smoothed scatter plot of the positions of these sensors and label five disconnected regions: UL, LL, UI, LI, and mouth cavity. Refer to the caption in Figure 1.a for the notation used here. The biggest of these regions is the mouth cavity composed of the region of the TT, TB, and TD. We set the highest points in the mouth cavity as the EMA's estimated upper wall. We add to this estimated upper wall the mean location of the velum sensor, VL.

The grid adaptation parameters are the width of the grid *d* and the increment β to the angle between the polar grids θ. We also estimate the inward/outward shift of the upper wall contour, $UPLT_{shift}$, referred to before as the palate height. This distance is added to O_{uwall} in Equation 1. We choose a range over which we vary each of the three parameters. For each combination we compute the average geometric distance between all the points on the estimated EMA and the adapted Maeda upper walls. These distances are shown in magenta in Figure 2. We choose the set of parameters with the least average distance. Note that the value of *d* reflects vocal tract stretching or compression with respect to the standard Maeda model and the value of β reflects oral tract tilt.

3.3. Lips Translation

The EMA lips sensors are placed outside the mouth on the tip of the lips [1]. This means that even when the lips are closed there is still a vertical gap between the two sensors. After adapting the upper wall, we estimate the minimum lips separation, Lip_{sep}, which is the hypothesized gap between the sensors when the lips are closed.

In Maeda's model, the outermost lower lip point has two degrees of freedom proportional to protrusion and separation. We map the four EMA measurements (UL, LL, UI, and LI) to a point LL_M in the vicinity of Maeda's outermost lower lip point. The protrusion is defined as the horizontal distance between the UI and the UL or the LL, whichever is smaller. The separation is defined as the vertical distance between UL and LL minus the estimated Lip_{sep}. For a demonstration of this translation, refer to Figure 3.

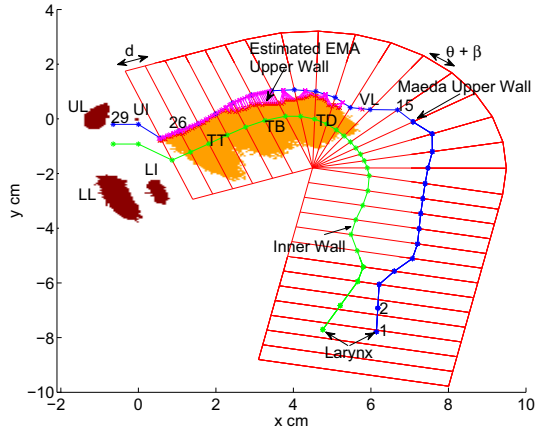


Fig. 2. Vocal tract adaptation showing the smoothed scatter plot of distribution of the EMA data (orange and brown) and the superimposed Maeda model (red grid lines). The green contour is for the steady state Maeda inner wall shape and the blue contour is the adapted Maeda upper wall.

4. VOCAL TRACT MODEL MATCHING

In [5], an approach for mapping EMA data to Maeda parameters has been described for the purpose of speech recognition. It uses a heuristic mapping from EMA directly to Maeda parameters without actually using the model. For example, p_5, is simply the normalized distance between the UI and LI.

The work in this paper describes a more principled approach that searches for the best fit of the EMA data to the adapted Maeda's vocal tract contours. We use a uniform codebook of Maeda parameters that represents different vocal tract shapes. For each frame of EMA data, we search for the best geometric fit. The best fit of the tongue and lip contours found for each frame of EMA data is then used in articulatory speech synthesis.

4.1. Codebook Design

We create a uniform codebook composed of 164K codewords, where p_1 to p_6 take values of {-3, -2.25, -1.5, -0.75, 0, 0.75, 1.5, 2.25, 3}. p_7 is set to zero since we do not have EMA data to estimate the larynx position. Some of the shapes in this codebook have constriction in the larynx region. We remove the shapes with an area less that 2cm^2 in the larynx region and are left with 16,850 codes.

4.2. Searching Vocal Tract Shapes

For each codeword, we compute the vocal tract profile and project it onto the adapted semi-polar coordinate space. For the EMA data, we first translate the UL, LL, UI, and LI to the LL$_M$ point as described in Subsection 3.3 and compute the distance from this point to the outermost lower lip point in the lower contour provided by the given codeword. Thus, we compute the first distance pertaining to the lips. Then for the TT, TB, and TD EMA points, we first find the grid section number in which each of these points falls. We then compute the distance of the EMA point to the segment of the lower vocal tract contour that falls within this grid section. Thus we find the three other distances. The overall geometric distance between the given frame of EMA data and the vocal tract contour of the given codeword is the mean of the above four distances. We choose the codeword that yields the least distance.

5. SYNTHESIS MODEL

Once we find the best matching vocal tract shape, we convert it to areas and lengths of the tubes forming the sections of the vocal tract. We follow Maeda's approach in computing the effective length and area of each tube bounded within the upper and lower contours and the grid lines. We then feed these areas and lengths to an articulatory speech synthesizer. We use the Sondhi and Schroeter [8] model[1] which uses the chain matrices approach to derive the overall transfer function of the vocal tract.

We replace the source modeling of Sondhi and Schroeter that uses the two-mass model of vocal cords developed by Ishizaka and Flanagan [10] with a modified version. The new approximation decouples the vocal tract from the glottis. For the transfer function, it uses Sondhi and Schroeter entire vocal tract transfer function, including the nasal tract. For generating the source signal, we use Rosenberg glottal pulse model [11] for voiced frames and random noise for unvoiced frames. We extract the energy and pitch from the original speech signal and use them in generating the source signal. This approach improves the synthesis quality and is faster than our previous approach [5].

5.1. Velum Location and Nasal Tract Opening Area

The Sondhi and Schroeter model also allows for nasal tract coupling to the vocal tract by adjusting the velum opening area. The location of the velum VL$_{loc}$ is set after adaptation to the grid section number that precedes (counting from the glottis to the lips) the one which contains the mean of VL. This is because the velum sensor is placed on the soft palate [1]. The velum opening area is estimated from the ordinate of the velum sensor VL$_y$. For each utterance, the nasal tract is opened proportional to how much the value of VL$_y$ is below its mean over the utterance.

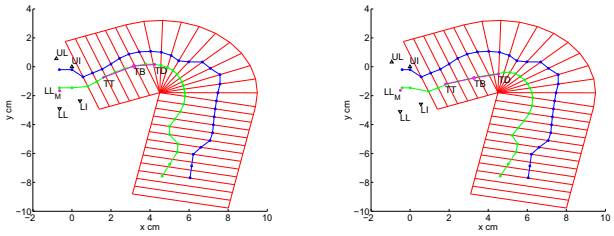
6. EXPERIMENTS

MOCHA contains EMA measurements and the corresponding acoustic speech signal for 10 speakers reading 460 TIMIT sentences. In this paper, we use the EMA data from speaker "msak0". We use all the EMA data available from this speaker to geometrically adapt Maeda's model. We use EMA data of 102 utterances to perform the synthesis and compare to the corresponding real speech. For each frame in the utterance we synthesize speech following the approach described in Sections 4 and 5. Then we extract Mel-cepstra coefficients (MFCC) from the synthesized and the real speech respectively and compute the MCD between them for each frame as in [5]. When computing the average MCD, we include the MCD of the frames at the onset, middle, and offset of phones. Phonetic segmentation has been automatically extracted beforehand.

6.1. Adaptation Results

As described in Subsection 3.2, we vary *d*, β, and $UPLT_{shift}$ until the projected Maeda upper wall best matches the estimated EMA upper wall. Before adaptation the average distance between the two contours is 0.65cm and the values of *d*, β, and $UPLT_{shift}$ are {0.5cm, 0rad, 0cm}. This distance is the average of the distances from each point on the EMA upper wall (red) to the Maeda wall (blue). These distances are shown in Figure 2 in magenta. The average distance between the two contours after adaptation it is 0.23cm

[1] We modify the implementation of Maeda's model and the Sondhi and Schroeter model provided with the articulatory synthesis package developed by Riegelsberger[9].



(a) 'II' (b) '@@'

Fig. 3. Search results for two EMA frames for 'II' in "Seesaw = /S-II-S-OO/" and '@@' in "Thirty = /TH-@@-T-II/". The EMA points used are LL_M, TT, TB, and TD shown in magenta. The resulting inner and upper walls of the matched shapes are in green and blue respectively.

and the values of d, β, and $UPLT_{shift}$ are $\{$0.55cm, -0.0033rad, 0.4cm$\}$. This means that the length of the vocal tract is extended by 10% and that the upper wall is shifted inwards by 0.4cm. These numbers make sense since the Maeda model is based on images from two female speakers and the MOCHA speaker "msak0" is a male speaker. In addition a close match is attained between the two upper wall contours. Note also that the mean of the velum sensor locations, VL, almost falls on the adapted Maeda upper wall.

6.2. Search Results

Notice in Figure 2 that the minimum separation between the lips, Lip_{sep}, is estimated to be 2cm. This measure is used to first translate the lips to the LL_M point. Figure 3 shows results of the search for the vocal tract shapes of the EMA data belonging to two frames in the middle of phones $\{$'II', '@@'$\}$ in the words "Seesaw = /S-II-S-OO/" and "Thirty = /TH-@@-T-II/". It is clear that the resulting vocal tract shapes fit well the projected EMA data and reflect the articulatory characteristics of the two phones. Phone 'II' is a high front vowel and phone '@@' is mid vowel.

6.3. Synthesis Results

We estimate the nasal tract opening area from the EMA measurements of the velum sensor VL. As shown in Figure 2, VL falls in grid section 16. We set VL_{loc} to 15 and estimate the opening area as described in Subsection 5.1. We compute the average MCD for frames from vowels, fricatives, nasals, and all the phones together. The baseline is the synthesis technique based on the EMA to Maeda parameters mapping described in [5] without adapting Maeda's model.

For the adapted vocal tract experiment, we achieve 9.77% relative reduction over baseline in MCD for vowels, 1.67% for fricatives, 5.84% for nasals, and 5.19% for all the phones together. Table 1 presents the results. Figure 4 shows the spectrogram of the original speech and the speech synthesized from the EMA measurements for an utterance in MOCHA using the adapted model. The corresponding audio example is available at http://www.cs.cmu.edu/%7Eziada/papers/interspeech09.

7. CONCLUSION AND FUTURE WORK

We presented a principled approach for mapping EMA data to vocal tract shapes for the task of speech synthesis by a physical model of the vocal tract. We relied solely on the EMA data to adapt Maeda's vocal tract model to a new speaker in the MOCHA database. We presented a way for searching for the best fitting vocal tract contours.

Table 1. MCD results: the absolute and relative differences are between the baseline experiment without adaptation and the adapted vocal tract approach developed in this paper.

MCD Results	Vowels	Fricatives	Nasals	Total Frames
Frame Count	3012	1296	846	8634
No Adaptation	8.40	7.85	8.75	8.52
Adapted Vocal Tract	7.58	7.72	8.24	8.08
Absolute Difference	0.82	0.13	0.51	0.44
Relative Difference (%)	9.77	1.67	5.84	5.19

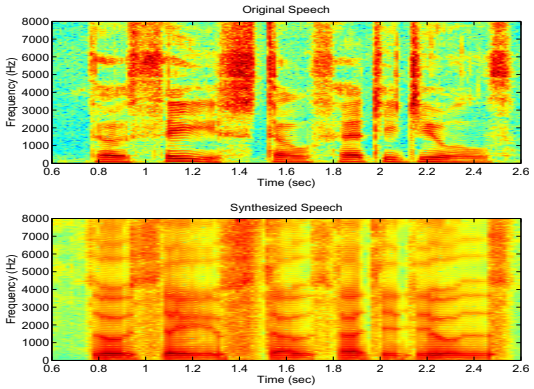


Fig. 4. Spectrogram of the original and synthesized speech for the utterance: "Those thieves stole thirty jewels".

Experiments showed improvement in synthesis over the baseline approach we adopted. In the future, we would like to use the ElectroPalatoGraph (EPG) data provided in MOCHA to improve modeling of fricatives and the constriction location.

8. ACKNOWLEDGMENT

We would like to thank Sankaran Panchapagesan for the discussions on this research, Bent Schmidt-Nielsen and MERL for the partial support of this work.

9. REFERENCES

[1] A. Wrench, "A new resource for production modeling in speech technology," in *Workshop on Innovations in speech processing*, Stratford-upon-Avon, UK, 2001.

[2] T. Toda, A. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA. USA, 2004, pp. 31–36.

[3] A. Toth and A. Black, "Using articulatory position data in voice transformation," in *ISCA SSW6*, Bonn Germany, 2007.

[4] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Modellling*. W.J. Hardcastle and A. Marchal, 1990, pp. 131–149.

[5] Z. Al Bawab, B. Raj, and R. M. Stern, "Analysis-by-synthesis features for speech recognition," in *ICASSP*, Las Vegas, Nevada. USA, April 2008.

[6] R. McGowan and S. Cushing, "Vocal tract normalization for midsagittal articulatory recovery with analysis-by-synthesis," *J. Acoust. Soc. Am.*, vol. 106, Issue 2, pp. 1090–1105, August 1999.

[7] B. Mathieu and Y. Laprie, "Adaptation of maeda's model for acoustic to articulatory inversion," in *Eurospeech*, Rhodes, Greece, 1997, pp. 2015–2018.

[8] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transac. ASSP*, vol. 35, pp. 955–967, July 1987.

[9] E. L. Riegelsberger, *The Acoustic-to-Articulatory Mapping of Voiced and Fricated Speech*, Ph.D. thesis, The Ohio State University, 1997.

[10] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Sys Tech J*, vol. 51, no. 6, pp. 1233–1268, 1972.

[11] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am*, vol. 49 2, pp. 583590, 1971.