MAXIMUM A-POSTERIORI ESTIMATION OF MISSING SAMPLES WITH CONTINUITY CONSTRAINT IN ELECTROMAGNETIC ARTICULOGRAPHY DATA

Sujith P

Electrical Communication Engineering
Indian Institute of Science (IISc)
Bangalore, India
sujith@ece.iisc.ernet.in

Prasanta Kumar Ghosh

Electrical Engineering
Indian Institute of Science (IISc)
Bangalore, India
prasantg@ee.iisc.ernet.in

ABSTRACT

Electromagnetic Articulography (EMA) technique is used to record the kinematics of different articulators while one speaks. EMA data often contains missing segments due to sensor failure. In this work, we propose a maximum a-posteriori (MAP) estimation with continuity constraint to recover the missing samples in the articulatory trajectories recorded using EMA. In this approach, we combine the benefits of statistical MAP estimation as well as the temporal continuity of the articulatory trajectories. Experiments on articulatory corpus using different missing segment durations show that the proposed continuity constraint results in a 30% reduction in average root mean square error in estimation over statistical estimation of missing segments without any continuity constraint.

Index Terms— EMA data, Missing sample estimation, Gaussian mixture model, Continuity constraint

1. INTRODUCTION

Electromagnetic Articulography (EMA) [1] provides the kinematics of sensors placed on articulator flesh points in the mid-sagittal plane. This is widely used for research in the articulation of speech, which offers an alternate representation to that in the acoustic domain. Other popular methods for articulatory movement recording are X-ray [2], X-ray microbeam [3], Magnetic Resonance Imaging (MRI) [4] and Ultrasound [5]. EMA data [6] consists of the X and Y coordinates of sensors placed on various articulators including upper lip, lower lip, jaw, tongue tip, tongue body, tongue dorsum and velum at a rate of 500Hz. The recorded data using EMA may contain missing segments because of sensor failure and sensor detachment [6]. However the recording of the articulatory movements using EMA is expensive in the sense that specialist facilities and expertise are typically required. The recording process itself can be tricky, with practical complication that may also increase burden on the subject [7]. Hence missing samples estimation technique [8] could be useful for reconstructing the missing samples rather than discarding the entire data containing missing segments.

In a previous work on missing samples estimation, Qin et al. [8] used minimum mean squared error (MMSE) criterion to predict the missing sample in each frame separately. Q. Fang et al. [9] used dynamical feature in addition to the articulatory feature to incorporate the temporal information. However they did not consider the fact that the trajectories formed by the movements of the articulators are smooth [10]. Lack of smoothness constraint could cause discontinuities at the boundaries between the missing and known segments as well as within the missing segments. Articulatory trajectories need to be slowly varying within the missing segments and also continuous at the boundaries between the missing segments and the known segments. This means that the reconstructed segments and the known segments should form a continuous trajectory. In contrast to MMSE, in this work we propose maximum a-posteriori estimation with continuity constraint (MAPC), where we constrain the estimates of the articulatory trajectories to be low-pass in nature. According to Toda et al. [11] “MMSE-based mapping is not appropriate for multiple probability density distributions because it ignores the covariances of the individual distributions even when they are different from each other”. In MAP estimation, the covariance matrices are used as weights in the weighted sum of the means of individual distributions unlike MMSE based estimation. Covariance matrices are regarded as confidence measure for the means from individual mixture components [11]. Thus MAP makes use of the statistical benefit in MAP estimation as well as the temporal continuity of the articulatory trajectory. We perform missing sample estimation on articulatory database and we find that, on average, MAPC achieves ~30% reduction in RMSE in estimation compared to MMSE.

2. MAXIMUM A-POSTERIORI ESTIMATION WITH CONTINUITY CONSTRAINT

We perform maximum a-posteriori (MAP) estimation of missing samples subject to the constraint that the energy of the high frequency contents of the articulatory trajectory after missing sample estimation is low; this ensures that the trajectory varies slowly in time. Let $x_n$ be a D dimensional articulatory feature vector, $K_{li}$ is the set of indices of known samples and $M_{li}$ is the set of indices of missing samples at the frame $n$. The estimator is a function that maps from known variables $x^m_{0:n}$ to missing variable $x^m_{n}$, where $m \in M_{li}$, $p$ is $K_{li}$ and $|K_{li}| = |M_{li}| = D$.

Let the g-th articulatory trajectory of length $N + 1$ be denoted by $X^q = (x^q_0, \ldots, x^q_{N_1}, \ldots, x^q_{N_q})$, $q \in M_{li}$. Let the samples from the frame $N_1$ to $N_2$ are missing and represented by $X^{q,m} = (x^{q,m}_{N_1}, \ldots, x^{q,m}_{N_2})$, where $x^{q,m}_n = x^q_n$, $N_1 \leq n \leq N_2$ i.e., the super script ‘m’ is used to indicate missing sample. Let $X^p = (x^p_0, \ldots, x^p_{N_2-1}, x^p_{N_2})$ be the set of vectors comprising samples of the known articulatory trajectories from the frame $N_1$ to $N_2$. The objective is to find an estimate of $X^{q,m}$ that maximizes the joint a-posteriori probability function $f(X^{q,m}) = p(x^{q,m}_0, \ldots, x^{q,m}_{N_2}|x^p_0, \ldots, x^p_{N_2})$ while minimizing the energy of the high frequency components in the estimated trajectory. $p(.)$ denotes probability density function. The energy of the high-frequency components can be expressed as $\sum_{n=0}^{\infty} y_n^2$, where

$q_i$ is the index of one of the articulators which contain at least one missing segment in a given utterance.
where $\lambda$ is a positive number and $\mathbf{X}^{q,m}$ is the set of optimization variables. We use a negative sign before the second term in eqn (5) because the first term in the objective function $f(\mathbf{X}^{q,m}, \mathbf{Z})$ has to be maximized while the second term has to be minimized. The variable $Z_{nr}$ in eqn (5) is also unknown, so we use the expectation maximization (EM) algorithm [13] for maximization. The objective function $f(\mathbf{X}^{q,m}, \mathbf{Z})$ in $(t+1)$-th iteration becomes $f(\mathbf{X}^{q,m}, \mathbf{Z}_{t}, \mathbf{X}^{q,m}_{t})$, where $\mathbf{X}^{q,m}_{t} \in \mathbb{R}^{N_{r} - N_{i} + 1}$ is a vector whose elements $x_{n,m}^{q}$ are the estimates of the samples of the missing segment $\mathbf{X}^{q,m}$ at the $t$-th iteration of the EM algorithm. In the first iteration of EM, we initialize $\mathbf{X}^{q,m}_{0}$ using the MMSE estimate and in subsequent iterations we use the estimate from previous iteration.

Below we describe the steps involved in the estimation process using the EM algorithm at the $(t+1)$-th iteration after discarding the terms that are independent of the optimization variables in eqn (5).

**Expectation step:**

$$f(\mathbf{X}^{q,m}) = E_{\mathbf{Z}}[f(\mathbf{X}^{q,m}, \mathbf{Z}, \mathbf{X}^{q,m})] = f_{1}(\mathbf{X}^{q,m}) + f_{2}(\mathbf{X}^{q,m})$$

$$= \sum_{n=N_{1}}^{N_{2}} \sum_{r=1}^{M} Z_{nr} \left( \log(\sigma_{nr}) - \frac{(x_{n,m}^{q,r} - \mu_{nr})^{2}}{2\sigma_{nr}^{2}} - \log(\sqrt{2\pi\sigma_{nr}}) \right) - \lambda \left( \sum_{n=0}^{N_{1} - 1} y_{n}^{2} + \sum_{n=N_{1}}^{\infty} y_{n}^{2} \right), \quad (5)$$

where $\lambda$ is a positive number and $\mathbf{X}^{q,m}$ is the set of optimization variables. We use a negative sign before the second term in eqn (5) because the first term in the objective function $f(\mathbf{X}^{q,m}, \mathbf{Z})$ has to be maximized while the second term has to be minimized. The variable $Z_{nr}$ in eqn (5) is also unknown, so we use the expectation maximization (EM) algorithm [13] for maximization. The objective function $f(\mathbf{X}^{q,m}, \mathbf{Z})$ in $(t+1)$-th iteration becomes $f(\mathbf{X}^{q,m}, \mathbf{Z}_{t}, \mathbf{X}^{q,m}_{t})$, where $\mathbf{X}^{q,m}_{t} \in \mathbb{R}^{N_{r} - N_{i} + 1}$ is a vector whose elements $x_{n,m}^{q,t}$ are the estimates of the samples of the missing segment $\mathbf{X}^{q,m}$ at the $t$-th iteration of the EM algorithm. In the first iteration of EM, we initialize $\mathbf{X}^{q,m}_{0}$ using the MMSE estimate and in subsequent iterations we use the estimate from previous iteration.

**Maximization step:**

For maximization we need to satisfy the first order necessary condition – differentiate $f(\mathbf{X}^{q,m})$ and equate it to zero.

$$\frac{\partial f_{1}}{\partial x_{n,m}^{q}} = x_{n,m}^{q} - \sum_{r=1}^{M} \frac{\bar{\pi}_{nr}^{r}}{\sigma_{nr}^{2}} + \sum_{r=1}^{M} \frac{\bar{\pi}_{nr}^{r}}{\sigma_{n}^{2}} \quad (7)$$

$$\frac{\partial f_{2}}{\partial x_{n,m}^{q}} = -\lambda \sum_{n=N_{1}}^{N_{2}} \sum_{k=0}^{\infty} h_{k} x_{n-k}^{q} h_{n-i} = \lambda \sum_{i=1}^{N_{2}} a_{ij} x_{n,m}^{q} + b_{i}, \quad (8)$$

where $a_{ij} = R_{h_{k}(i-j)} \sum_{n=0}^{\infty} h_{n} h_{n-(i-j)}$, $b_{i} = \sum_{t=1}^{\infty} \sum_{k=0}^{\infty} \bar{x}_{n-k}^{q,h_{k} h_{n-i}} \sum_{i=0}^{\infty} R_{h_{k}}$, $x_{n}^{q}$ otherwise

$$\frac{\partial f}{\partial \sigma_{n}^{2}} = -\frac{\partial f_{1}}{\partial \sigma_{n}^{2}} + \frac{\partial f_{2}}{\partial \sigma_{n}^{2}} = 0, \quad N_{1} \leq i \leq N_{2}, \text{ we can form}$$

2Although we have used causal IIR filter in the present formulation, one can choose any FIR or IIR filter

$^3$Note that, as logarithm is a monotonic function, this does not change the solution of the optimization
a matrix-vector equation as follows:

\[
(A_1 + \lambda A_2)x_{t+1}^{q,m} + (B_1 + \lambda B_2) = 0,
\]

where matrices $A_1$ and $B_1$ come from eqn (7), $A_2$ and $B_2$ come from eqn (8).

Thus, $A_1$ and $A_2$ are square matrices and $(B_1 + \lambda B_2)$ is a column matrix. $A_2$ is the auto-correlation matrix of filter impulse response. Hence, $A_1$ is a Toeplitz and positive semidefinite matrix. $A_1 + \lambda A_2$ is a positive definite matrix and the inverse of $(A_1 + \lambda A_2)$ exists. From eqn (7) and (8) the Hessian is a negative definite matrix that gives the second order sufficient condition. Thus, from eqn (10), we obtain the estimate at the $(t+1)$-th iteration as follows:

\[
x_{t+1}^{q,m} = -(A_1 + \lambda A_2)^{-1}(B_1 + \lambda B_2)
\]

We use the following stopping criterion for the EM iterations:

\[
\frac{1}{N_2 - N_1 + 1} \|x_{t+1}^{q,m} - x_t^{q,m}\|_2 < 10^{-4}
\]

3. EXPERIMENTAL EVALUATION

3.1. DATASET

For the experiments, we use the Multichannel Articulatory (MOCHA) [6] database which contains 460 utterances each spoken by one male and one female subject. Preprocessing is done by following the steps outlined in [14] and thus the acoustic and articulatory data at a rate of 100 frames per second are obtained. We use 14 dimensional EMA feature vector in each frame denoted by ULx, LLx, JAWx, TTx, TBx, TDX, VELx, ULy, LLy, JAWy, TTy, TBy, Tdy, and VELy. Acoustic features are represented by 39 dimensional Mel-frequency cepstral coefficients (MFCCs) and are computed using an analysis window of 20 ms length with 10 ms shift.

3.2. EXPERIMENTAL SETUP

We follow the same experimental procedure used by Qin et al. [8] by reconstructing artificially blacked-out portion of one articulator’s trajectory, which is treated as the missing segment. The initial point of the missing segment is selected at random and we consider four different missing segment durations (MSD) 100ms, 200ms, 400ms and 800ms. We also black out both X and Y trajectories of one sensor simulating the failure of that sensor. Thus missing segments of two trajectories are estimated using remaining 12 sensor’s data. The joint distribution $p(x^{q,m}_n)$ is modeled using the training data with GMM with 8 components. We have not got significant performance gain by using higher number of mixture components. We consider a five-fold cross validation set-up using 368 sentences for training and remaining 92 as the test set in each fold. Here we compare the performances of the MMSE and the MAPC for missing sample estimation task. In addition to the known articulatory features we also used acoustic features in the respective frames to investigate if there is any performance benefit by using acoustic features. We refer to this scheme by MAPCA. We also estimate the missing samples using a linear function (LIN). In LIN the missing samples are estimated as the points on the line joining the two end points of the missing segment $x_{N_2-1}$ and $x_{N_2+1}$. The equation of the line is $x^{q,m}_n = \alpha n + x_{N_2+1} - \alpha (N_2 + 1)$, where $n$ is the frame index and\[\alpha = \frac{x_{N_2+1} - x_{N_2-1}}{N_2 - N_1 + 2}.
\]

![Fig. 1. RMSE (average ± one standard deviation) of estimation for different articulators for different MSD for male speaker.](image)

Experiments are conducted using a Chebyshev FIR\(^4\) high-pass filter of order $L$. We consider different values of the hyper-parameters $-f_c$, $L$, and $\lambda$. Experiments are repeated with the following parameter sets: $f_c = \{5\text{Hz}, 10\text{Hz}, 20\text{Hz}, 40\text{Hz}\}$, $L = \{20, 30, 60\}$ and $\lambda = \{1, 10^{-1}, 10^{-4}, 10^{-7}, 10^{-10}, 10^{-12}\}$. We report the result corresponding to the parameter set which yields the best performance on the test set, i.e., $L=30$, $f_c = 20\text{Hz}$ and $\lambda = 10^{-5}$.

3.3. RESULTS AND DISCUSSION

Fig. 1 and Fig. 2 show the reconstruction performances in terms of the root mean square error (RMSE) between the original and estimated segments for different MSD for male and female subjects respectively. The $x$-axis represents various MSD and the $y$-axis represents the RMSE in mm scale. From the figures we can see that in terms of average RMSE, MAPCA\(^5\) performs better than MMSE and the performance gain of MAPC decreases with increasing MSD. For example, the reductions in the average RMSE by using MAPC over MMSE for the male subject are 76.3%, 25.6%, 11.9% and

\(^4\)Although an IIR filter could be chosen as the high-pass filter, we have experimented with FIR filter in this work.

\(^5\)We have also performed ML estimation by maximizing $p(x^{q,m}_n | z^{q,m}_n)$ but the results are inferior to that of the MMSE and MAP estimation.
5.5% for the MSD of 100ms, 200ms, 400ms and 800ms respectively. In the case of female subject, these reductions are 78.9%, 26.9%, 12.9% and 6.2%. The performance of MMSE is almost constant for different MSD. This is because in MMSE estimation each sample of a trajectory is estimated independently from the other samples in that trajectory.

The performance of LIN decreases with increasing MSD. Considering all articulators of both subjects, the average values of RMSE in the case of LIN are 0.56, 1.22, 1.96 and 2.48 mm for MSD of 100ms, 200ms, 400ms and 800ms respectively. We can see that for small MSD, LIN is better than MMSE but MAPC is even better than LIN in most of the cases except for VELx and VELy. In the case of MAPC, average RMSE reduces by 52%, 20%, 39% and 46% (for four MSD) over LIN. Thus, overall MAPC is found to yield lower average RMSE compared to both the baseline (LIN) and the MMSE approach. MAPC is better than MAPC but the average reductions in RMSE are 0.3%, 2.5%, 2.8% and 2.9% for the MSD 100ms, 200ms, 400ms, and 800ms respectively. Thus there is no significant benefit by using acoustic features in addition to the known articulatory features for missing samples estimation task. We observe that MAP (i.e., MAP without continuity constraint) performs similar to MMSE. This is because the conditional density $p(x^{m}_{n}|x^{m}_{p})$ is found to be dominated by a single Gaussian in most of the frames resulting in similar estimates using MAP and MMSE. Since the performance of MAP and MMSE are almost similar, the benefit is coming from the continuity constraint.

4. CONCLUSIONS

We have proposed a maximum a-posteriori estimation with continuity constraint (MAPC) for the task of missing sample reconstruction in articulatory trajectory. We have found that adding continuity constraint improves the performance of the missing sample reconstruction. However, the benefit of using continuity constraint decreases with increasing missing segment length. When a single sensor data is missing, we have found that there is no additional reconstruction benefit by using acoustic features in addition to the articulatory data from remaining known sensors. The estimation performance varies across different articulators; this could be due to different correlation between missing articulator and the known articulators. MAPC approach could also be useful for missing value imputation in other applications such as gene expression data [15, 16] and feature reconstruction in automatic speech recognition [17, 18].

---

*Polynomial interpolation is also carried out but the results are similar to that of LIN method.*

---

MAP+low-pass filter is not considered because it does not guarantee the continuity at the boundary of known and missing segments.

5. REFERENCES


