# Formant Trajectories for Acoustic-to-Articulatory Inversion

*İ. Yücel Özbek.*[1]*,Mark Hasegawa-Johnson*[2]*, Mübeccel Demirekler*[1]

[1]EE Department, Middle East Technical University, Turkey
[2]ECE Department, University of Illinois at Urbana-Champaign, US
iozbek@metu.edu.tr,jhasegaw@uiuc.edu, demirek@metu.edu.tr

## Abstract

This work examines the utility of formant frequencies and their energies in acoustic-to-articulatory inversion. For this purpose, formant frequencies and formant spectral amplitudes are automatically estimated from audio, and are treated as observations for the purpose of estimating electromagnetic articulography (EMA) coil positions. A mixture Gaussian regression model with mel-frequency cepstral (MFCC) observations is modified by using formants and energies to either replace or augment the MFCC observation vector. The augmented observation results in 3.4% lower RMS error, and 2.7% higher correlation coefficient, than the baseline MFCC observation. Improvement is especially good for plosive consonants, possibly because formant tracking provides information about the acoustic resonances that would be otherwise unavailable during plosive closure and release.

**Index Terms**: acoustic-to-articulatory inversion, formant tracking, GMM regression

## 1. Introduction

Formant frequencies are the resonances (natural frequencies) of the vocal tract. As the articulators move, the vocal tract area function changes, and therefore the resonance frequencies of the vocal tract change. Hence, there is a close relation between position of articulators and formant frequencies. There are numerous studies in the literature in which formant frequencies are considered as acoustic data to estimate corresponding articulatory data [3, 9].

The aim of this study is to examine the usefulness of formant related acoustic features as inputs to a Gaussian-mixture-model regression (GMM) estimator of articulator positions. GMM regression has been demonstrated to successfully estimate the positions of receiver coils in Electromagnetic Articulography (EMA) recordings, using mel-frequency cepstral coefficients (MFCC) as the observation [4]. The utility of formants as an input to other articulatory estimation methods suggest the possibility that formant-based parameters may also be useful for GMM regression. This paper measures their utility.

The rest of the paper is organized as follows: Section 2 gives a summary of the GMM based non-linear regression method for articulatory inversion. Section 3 describes extraction formant related acoustic features. The experimental results are given in Section 4. Section 5 presents our conclusions and discussion.

## 2. Acoustic-to-articulatory inversion

GMM based nonlinear regression is used in acoustic to articulatory mapping [4]. The basic idea of this method is as follows. Let $\mathcal{Z}$, $\mathcal{Y}$ be two vectors from acoustic and articulatory spaces and let $\mathbf{g}(.)$ be an inverse mapping function defined as:

$$\mathcal{Y} = \mathbf{g}(\mathcal{Z}) \tag{1}$$

Acoustic-to-articulatory inversion methods look for an inverse mapping function $\mathbf{g}$ (.) to estimate articulatory vectors from given acoustic data. In a probabilistic framework, the inverse mapping function $\mathbf{g}$ (.) can be approximated if enough data pairs ($\mathcal{Z}_i$ , $\mathcal{Y}_i$) are available. Let $\hat{\mathbf{g}}$ (.) be an estimate of the true inverse mapping function $\mathbf{g}$ (.). Suppose that $\hat{\mathbf{g}}$ is selected in order to minimize the mean squared error of the articulatory estimate, thus

$$\hat{y} \triangleq \hat{\mathbf{g}}(\mathcal{Z}) = E(\mathcal{Y}|\mathcal{Z}) \tag{2}$$

Assume that $\mathcal{Z}$, $\mathcal{Y}$ are jointly distributed according to a mixture Gaussian probability density function. In that case, the joint distribution can be written as

$$f_{\mathcal{Y},\mathcal{Z}}(y,z) \triangleq \sum_{i=1}^{K} \pi_i \mathcal{N}(y,z;\mu^i,\Sigma^i) \tag{3}$$

where, $\mu^i = \begin{bmatrix} \mu^i_{\mathcal{Y}} \\ \mu^i_{\mathcal{Z}} \end{bmatrix}$, $\Sigma^i = \begin{bmatrix} \Sigma^i_{\mathcal{Y}\mathcal{Y}} & \Sigma^i_{\mathcal{Y}\mathcal{Z}} \\ \Sigma^i_{\mathcal{Z}\mathcal{Y}} & \Sigma^i_{\mathcal{Z}\mathcal{Z}} \end{bmatrix}$
$K$ is the number of mixture components, and the mixture weights $\pi_i$ satisfy $\sum_{i=1}^{K} \pi_i = 1$. The conditional expectation $E(\mathcal{Y}|\mathcal{Z})$ is then:

$$\hat{y} \triangleq E(\mathcal{Y}|\mathcal{Z}) = \sum_{i=1}^{K} \beta^i(z)(\Sigma^i_{\mathcal{Y}\mathcal{Z}}(\Sigma^i_{\mathcal{Z}\mathcal{Z}})^{-1}(z-\mu^i_{\mathcal{Z}}) + \mu^i_{\mathcal{Y}}) \tag{4}$$

$$\beta^i(z) = \frac{\pi_i \mathcal{N}(z;\mu^i_{\mathcal{Z}},\Sigma^i_{\mathcal{Z}\mathcal{Z}})}{\sum_{i=1}^{K} \pi_i \mathcal{N}(z;\mu^i_{\mathcal{Z}},\Sigma^i_{\mathcal{Z}\mathcal{Z}})} \tag{5}$$

Eq. 4 shows that, under the assumed mixture Gaussian distribution, $E(\mathcal{Y}|\mathcal{Z})$ is a weighted average of affine functions. The parameter set of the GMM, $\Theta = (\pi_i, \mu^i, \Sigma^i)$, may be estimated using the expectation maximization algorithm, as described in [4].

## 3. Extraction of formant frequencies and their energies

Formants are the resonant frequencies of the vocal tract. During vowels and glides, formant frequencies may be estimated by the poles of an autoregressive spectral estimator, though temporal smoothing improves the estimate; during obstruent consonants, formant frequencies must be interpolated using some type of

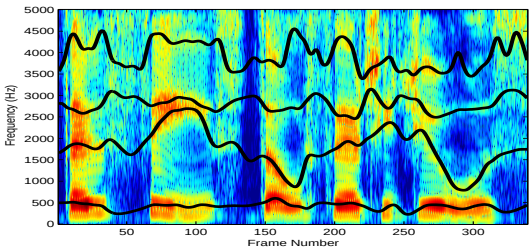


Figure 1: *The spectrogram of the utterance 'Those thieves stole thirty jewels' from fsew0-Mocha-TIMIT database. Estimated formant trajectories are superimposed.*

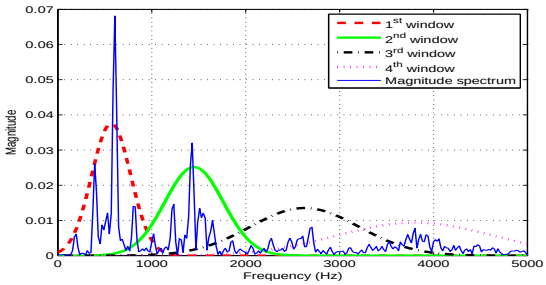


Figure 2: *Magnitude spectrum and Gaussian windows for 155'th frame of Fig-1. Corresponding four formant frequencies are F=[586, 1457, 2628, 3803] Hz.*

dynamic programming model. In this paper we use the formant tracker described in [11]. Our formant tracker consists of three stages, two of which are quite standard, and one of which is unusual but useful. In the first stage, frame-based formant candidates and their bandwidths are calculated by solving the denominator polynomial of the LPC filter. In second stage, formants are selected from among the candidates using a dynamic programming algorithm. In the third stage, formant trajectories are re-estimated using a Kalman smoother. LPC analysis is based on a spectral observation of 5kHz bandwidth (10kHz sampling frequency), using a 12th order autoregressive model. The output of the formant tracking algorithm for the one of the sentences from Mocha-Timit database is shown in Fig. 1. The energy associated with each formant frequency is calculated as follows. First, a magnitude spectrum is computed for each frame. Second, for each formant, Gaussian windows are generated in the spectrum domain. The mean and variance of each Gaussian are related to the associated formant frequency and bandwidth respectively. The bandwidths of the first four formants are assumed to be fixed at the values of BW= [90, 110, 170, 220] Hz. The means of the Gaussian windows vary in time, tracking the estimated formant frequencies. Third, the energy level associated with the $i$th formant, $E_i$, is computed by multiplying the magnitude spectrum $|X(f)|$ by the $i$th Gaussian window, $G_i(f)$, and summing over all frequencies:

$$E_i = \ln\left(\sum_{f=0}^{F_s} G_i(f)|X(f)|\right) \quad (6)$$

Fig. 2 shows the magnitude spectrum and corresponding Gaussian windows for the 155'th frame of the spectrogram given in Fig. 1.

## 4. Experiments

### 4.1. Experimental condition

In this work, we use the Timit-MOCHA database [1]. The acoustic data and EMA trajectories of one female talker (fsew0) are used; these data include 460 sentences. MFCC and formant related features were computed using a 36ms window with 18ms shift. The acoustic feature types used in this work are given in Table-1. The articulatory data are EMA trajectories, which are the X and Y coordinates of the lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum and velum. The articulatory data are normalized by suggested methods given in [2] and downsampled to match the 18ms shift rate. All models are tested using 10-fold cross-validation. For each fold, nine tenths of the data (414 sentences) are used for training and one tenth (46 sentences) for testing. The estimated EMA trajectories are calculated by using equation (4) and these estimated trajectories are smoothed by the low pass filter described in [2]. Cross-validation performance measures (RMS error and correlation coefficient) are computed as the average of all ten folds.

Table 1: *Acoustic feature types.*

| | |
|---|---|
| F | Four formant frequencies<br>F = [F1, F2, F3, F4] |
| E | Energy levels of four formants<br>E = [E1, E2,E3,E4] |
| M | Mel-frequency cepstral coefficients<br>(13 orders) M = [M1,...,M13] |
| X_$\Delta$, $\Delta\Delta$ | Combination of X and its time derivatives; velocity and acceleration components<br>X_$\Delta$, $\Delta\Delta$= [X, X_$\Delta$, X_$\Delta\Delta$]<br>(X can be any feature type or any combination<br>i.e. MF_$\Delta$, $\Delta\Delta$= [MF, MF_$\Delta$, MF_$\Delta\Delta$]) |

Algorithm performance is measured using the three performance measures (RMS error, normalized RMS error and correlation coefficient) described in [2, 10].

RMS error:

$$E^i_{RMS} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x^i_k - \hat{x}^i_k)^2}, i = 1, ..., m \quad (7)$$

where, $x^i_k$ and $\hat{x}^i_k$ are true and estimated position, respectively, of the $i$th articulator in the $k$th frame.

Normalized RMS error:

$$E^i_{NRMS} = \frac{E^i_{RMS}}{\sigma_i}, i = 1, ..., m \quad (8)$$

where, $\sigma_i$ is the standard deviation of $x^i$.

Correlation coefficient:

$$\rho^i_{x,\hat{x}} = \frac{\sum_{i=1}^{N}(x^i_k - \bar{x}^i_k)(\hat{x}^i_k - \bar{\hat{x}}^i_k)}{\sqrt{\sum_{i=1}^{N}(x^i_k - \bar{x}^i_k)^2}\sqrt{\sum_{i=1}^{N}(\hat{x}^i_k - \bar{\hat{x}}^i_k)^2}}, i = 1, ..., m \quad (9)$$

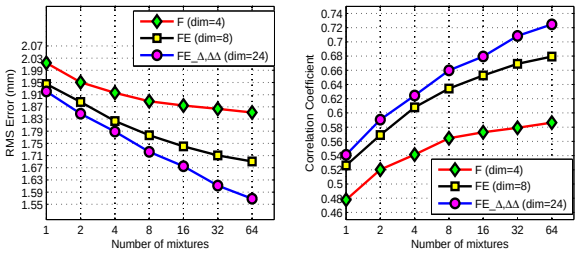


Figure 3: *RMS error and correlation coefficient as a function of the number of mixture components using formant related acoustic features given Table-1.*

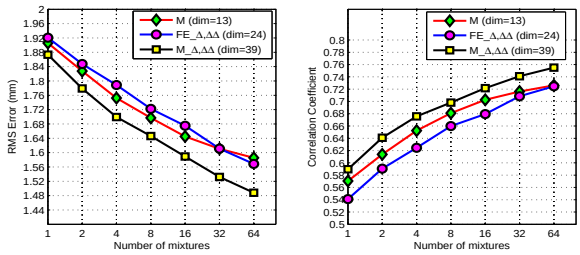


Figure 4: *RMS error and correlation coefficient as a function of the number of mixture components using formant features and MFCC.*

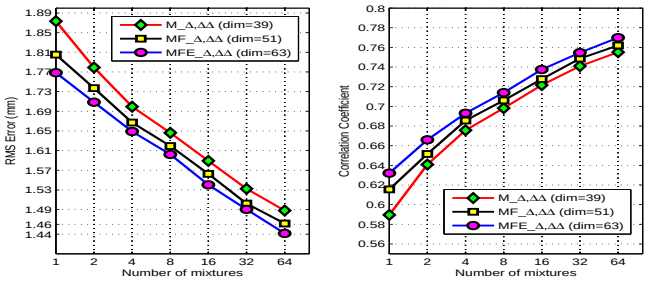


Figure 5: *RMS error and correlation coefficient using combination of formants related features and MFCC.*

where, $\bar{x}^i$ and $\bar{\hat{x}}^i$ are the average position of true and estimated $i^{th}$ articulator respectively.

In order to determine whether or not articulatory inversion using formant parameters significantly outperforms articulatory inversion without formant parameters, the following test was performed. Two hypothesis are used for this purpose: $H_0$ and $H_1$. The null hypothesis $H_0$ states that the RMS error of the regression model observing both formant parameters and MFCC is no different from the RMS error of the regression model observing only MFCCs; the test hypothesis $H_1$ states that the RMS errors differ. Thus:

$$H_0 : e = J - J^F \leq 0$$
$$H_1 : e = J - J^F > 0 \qquad (10)$$

where, $J$ is the RMS error without formant related features and $J^F$ is the RMS error with formant related features. As described in [5], we reject the null hypothesis if

$$Z = \frac{\bar{e}}{\frac{\sigma_{\bar{e}}}{\sqrt{K}}} > t_0(\alpha) \qquad (11)$$

where, $\bar{e} = \frac{1}{K}\sum_{i=1}^{K} e_i$, $\sigma_{\bar{e}} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(e_i - \bar{e})^2}$ and $t_0(\alpha)$ is the threshold based on the upper tail of the normal density with significance level $\alpha$; for $\alpha = 0.01$, $t_0 = 2.33$. In order to validate the assumption of independent trials, each sentence is treated as a trial, rather than each frame; thus $e_i$ is the average RMS for the $i$th sentence. There are $K = 460$ sentences in the ten-fold cross-validation test. Significance tests for correlation coefficients are performed using a similar procedure.

### 4.2. Experimental results

The acoustic-to-articulatory inversion experiment using only formant related acoustic features can be seen in Fig. 3. In this figure RMS error and correlation coefficient are calculated for different mixture Gaussian PDFs with 1 to 64 components. It can be observed that combination of formant related features and their velocity and acceleration component gives best performance for the 64-Gaussian regression. The resulting lowest RMS error is about 1.56 mm, and the highest correlation coefficient is about 0.72.

The comparison of formant related acoustic features and MFCC can be examined in Fig. 4. In general, MFCC give better results than formant related features; this is as reported in automatic speech recognition applications. This figure also shows that velocity and acceleration components improve the accuracy of MFCC-based articulatory inversion.

The combination of MFCC and formant related acoustic features in acoustic-to articulatory inversion is tabulated in Fig. 5. The RMS error is about 1.49 mm for M_$\Delta$, $\Delta\Delta$ with 64 Gaussian mixtures. Using only combination of MFCC and formant frequencies reduces RMS error to 1.46 mm. Combination of MFCC, formants and formant energies reduces RMS error to about 1.44 mm. Hence, overall RMS error reduction is about 3.35%. Correlation coefficient increases from 0.75 to 0.77, a 2.66% relative improvement.

Fig. 6 provides more details regarding the utility of formant related acoustic features in inversion. All results in this figure use a 64-Gaussian regression. The abscissa distinguishes different articulators. As an example, Normalized RMS error for Y axis of upper lip (uly) reduced from 0.736mm to 0.7mm, a 4.2% relative error reduction (left side of Fig. 6). Similarly, correlation improvements and corresponding percentages are given on the right side of the same figure.

Fig. 7 measures the significance of the normalized RMS error reductions and correlation coefficient improvements shown in Fig. 6. This figure shows that RMS error reduction and correlation improvement for each articulators are significant at the $\alpha = 0.01$ level of significance.

The experimental results also show formant related acoustic features are especially useful for plosive and fricative sounds, as well as vowel sounds (Table-2 and Fig. 8). As an example, RMS error reduction and correlation improvement are about 4% and 3.4% for plosive sounds, respectively.

An example of true and estimated trajectories for the y-coordinates of the tongue body is shown in Fig. 9 for a MOCHA utterance.

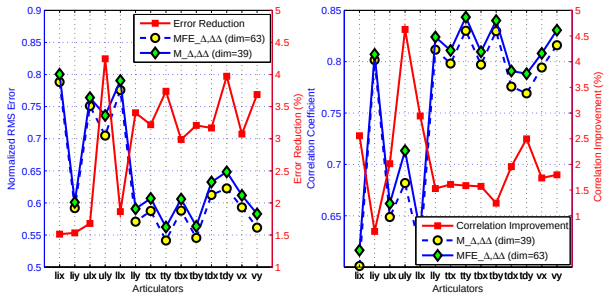


Figure 6: *Normalized RMS error reduction and correlation coefficient improvement for each articulator in detail. The number of Gaussian mixture is 64. li, ul,ll, tt, tb, td and v show lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum and velum, respectively. *x and *y in each articulator show X and Y coordinates, respectively*

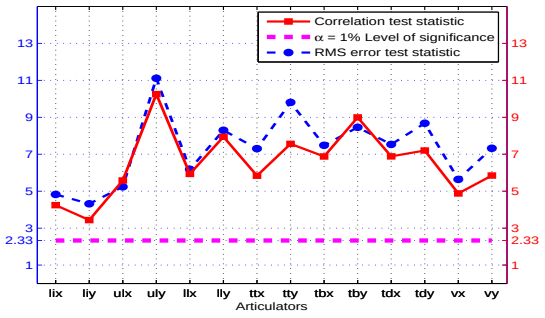


Figure 7: *Significance tests for normalized RMS error reductions and correlation improvements given in Fig. 6. Abbreviations related to the names of the articulators are explained in Fig. 6*

## 5. Discussion and conclusion

In this study, the usefulness of formant frequencies and corresponding energies in acoustic-to articulatory inversion are examined. It is observed that combination of MFCC and formant related features as acoustic features gives better results than using only MFCC. The average RMS error reduction is about 3.4%, and correlation improves by 2.7%; both improvements are statistically significant at the $\alpha = 0.01$ level of significance. Formant features are especially useful for articulatory inversion during plosives and fricatives; during plosive phonemes, RMS error reduction and correlation improvement are about 4% and 3.4%, respectively.

## 6. Acknowledgment

We would like to thank The Scientific and Technological Research Council of Turkey (TUBITAK) for its financial support

## 7. References

[1] A. Wrench, http://www.cstr.ed.ac.uk/artic/mocha.html, Queen Margaret University College, 1999.

[2] K. Richmond. Estimating Articulatory Parameters from the Speech Signal. PhD thesis, The Center for Speech Technology Research, Edinburgh, UK, 2002.

[3] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," JASA, vol. 118, no. 1, pp. 444-460, 2005.

[4] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," Speech Communication, vol. 50, pp. 215-227, 2008.

Table 2: *RMS error and Correlation coefficient for broad phonetic classes (vowel,approximate,nasal,plosive and fricative), using a 64-Gaussian regression.*

| | RMS eror (mm) | | | Correlation coefficient | | |
|---|---|---|---|---|---|---|
| Class | M_ Δ, ΔΔ | MFE_ Δ, ΔΔ | red (%) | M_ Δ, ΔΔ | MFE_ Δ, ΔΔ | imp (%) |
| Vowel | 1.424 | 1.379 | 3.2 | 0.759 | 0.775 | 2.1 |
| App. | 1.557 | 1.521 | 2.3 | 0.706 | 0.717 | 1.6 |
| Nasal | 1.568 | 1.527 | 2.6 | 0.645 | 0.659 | 2.2 |
| Plos. | 1.603 | 1.539 | 4 | 0.679 | 0.702 | 3.4 |
| Fric. | 1.406 | 1.369 | 2.6 | 0.638 | 0.658 | 3.1 |

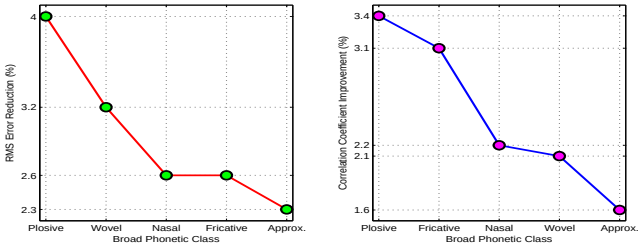


Figure 8: *Articulatory inversion RMS error reduction and Correlation coefficient improvement using formant related acoustic features for each broad phonetic class, using a 64-Gaussian regression.*

[5] Y. Bar-Shalom, X.R. Li, Estimation and Tracking: Principles, Techniques and Software,Artech House, Inc., 1993.

[6] O. Engwall, "Introducing visual cues in acoustic-toarticulatory inversion," in Interspeech, 2005, pp. 3205-3208.

[7] Asterios Toutios, Konstantinos Margaritis: "Contribution to Statistical Acoustic-to-EMAMapping",(Eusipco-2008),

[8] Qin, C. and Carreira-Perpin, M. . (2007) "A comparison of acoustic features for articulatory inversion".Interspeech 2007.

[9] Schroeter, J., Sondhi, M.M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trans.Sp.Au.Process.2,133-150.

[10] Katsamanis, A. and Papandreou, G. and Maragos, P,"Face Active Appearance Modeling and Speech Acoustic Information to Recover Articulation" IEEE Trans. Speech and Audio Proc., 17(3):411-422, 2009

[11] Özbek İ. Yücel, Mübeccel Demirekler "Vocal Tract Resonances Tracking Based on Voiced and Unvoiced Speech Classification Using Dynamic Programming and Fixed Interval Kalman Smoother" ICASSP-2008 Lasvegas, USA

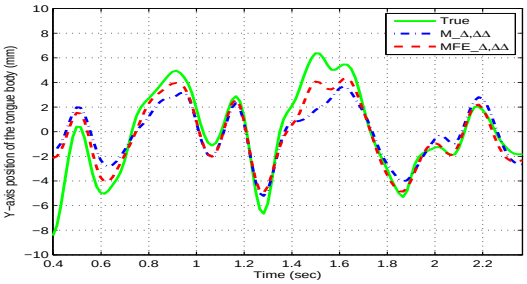


Figure 9: *Tongue body y-axis true and estimated trajectories as an example 'They all enjoy ice cream sundaes' from fsew0-Mocha-TIMIT database*