ACOUSTIC-TO-ARTICULATORY INVERSE MAPPING
USING AN HMM-BASED SPEECH PRODUCTION MODEL

Sadao HIROYA and Masaaki HONDA

NTT Communication Science Laboratories, NTT Corporation / CREST, JST
3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa, 243-0198 Japan
e-mail: {hiroya, hon}@idea.brl.ntt.co.jp

ABSTRACT
We present a method that determines articulatory movements from speech acoustics using an HMM (Hidden Markov Model)-based speech production model. The model statistically generates speech acoustics and articulatory movements from a given phonemic string. It consists of HMMs of articulatory movements for each phoneme and an articulatory-to-acoustic mapping for each HMM state. For a given speech acoustics, the maximum a posteriori probability estimate of the articulatory parameters of the statistical model is presented. The method’s performance on sentences was evaluated by comparing the estimated articulatory parameters with observed parameters. The average rms error of the estimated articulatory parameters was 1.79 mm with phonemic information and 2.16 mm without phonemic information in an utterance.

1. INTRODUCTION
The inverse problem for determination of articulatory movements from speech acoustics is characterized by one-to-many mapping [1]. In uniquely determining articulatory movements from speech acoustics, dynamic constraints on articulatory movements have been often used. In previous works, Schroeter and Sondhi [2][3], Hogden et al.[4], and Suzuki et al.[5] presented methods of an articulatory-acoustic codebook search with dynamic constraints on the continuity of the articulatory codevector sequence. Okadome et al.[6] incorporated the phonemic information of an utterance. Dusan and Deng [7] presented a method based on an extended Kalman filter.

However, previous works have not extensively taken into account the dynamic features of articulatory movements. To take into account the dynamic constraints beyond the continuity constraints, Hiroya and Honda [8] proposed a method of inverse mapping using an HMM-based speech production model.

In this paper, we present a method of estimating the maximum a posteriori probability of the articulatory movements for a given speech acoustics using an HMM-based speech production model. This model consists of HMMs of articulatory movements for each phoneme and an articulatory-to-acoustic mapping that transforms the articulatory parameters into acoustic parameters for each HMM state. The model was constructed by using simultaneously obtained articulatory and acoustic data for sentence utterances, which was collected by using an electro-magnetic articulographic system. Using the speech production model, HMM state sequence is determined by finding the maximum likelihood estimate of a given speech acoustics. Then articulatory parameters are determined by finding the maximum a posteriori estimate of articulatory parameters for a given speech acoustics and the HMM state sequence. This method was evaluated by comparing the estimated articulatory movements with observed movements.

2. HMM-BASED SPEECH PRODUCTION MODEL

The HMM-based speech production model is shown in Fig. 1. The model consists of HMMs which represent the articulatory movements, called articulatory HMM, and an
articulatory-to-acoustic mapping. The articulatory HMM has multiple states for each phoneme and generates articulatory parameter vector in a probabilistic form for a given phoneme sequence. For a given articulatory parameter vector, the articulatory-to-acoustic mapping generates an acoustic parameter vector in a probabilistic form for each HMM state. Denoting an articulatory parameter vector as \( x \) and an acoustic parameter vector as \( y \), the output probability of an acoustic parameter vector in the speech production model is represented as

\[
P(y|\lambda) = \sum_q \int P(y|x, q, \lambda)P(x|q, \lambda)P(q|\lambda)dx. \tag{1}
\]

Here, \( \lambda \) is phoneme-specific models of HMM and \( q \) is HMM state sequence. \( P(y|x, q, \lambda) \) is the occurrence probability of an acoustic parameter vector for a given articulatory parameter vector, and \( P(x|q, \lambda) \) is the output probability of a given HMM state. Assuming these probabilities to Gaussian distributions and the linear function \( y = Ax + b \), which approximates the articulatory-to-acoustic mapping \( y = f(x) \),

\[
P(y|x, q, \lambda) = \frac{1}{(2\pi)^{N/2} |\sigma_x|^{1/2}} \times \exp \left[ -\frac{1}{2}(y - Ax - b)^T \sigma_x^{-1}(y - Ax - b) \right] \tag{2}
\]

\[
P(x|q, \lambda) = \frac{1}{(2\pi)^{M/2} |\sigma_y|^{1/2}} \times \exp \left[ -\frac{1}{2}(x - \bar{x})^T \sigma_y^{-1}(x - \bar{x}) \right], \tag{3}
\]

where \( M \) and \( N \) are the dimensions of \( x \) and \( y \), \( \bar{x} \) and \( \sigma_x \) are the mean and covariance of \( x \), and \( \sigma_y \) is the covariance of the error \( w \) in the linear approximation of the articulatory-to-acoustic mapping for each HMM state.

3. DETERMINATION PROCEDURE

3.1. Estimation of articulatory parameter vector

For a given acoustic parameter vector \( y \) and an HMM state sequence \( q \), an articulatory parameter vector is determined by maximizing a posteriori probability

\[
P(x|y, q, \lambda) = \frac{P(y|x, q, \lambda)P(x|q, \lambda)}{P(y|q, \lambda)}. \tag{4}
\]

Then the maximum a posteriori estimate \( \hat{x} \) of an articulatory parameter vector is given by

\[
\hat{x} = (\sigma_x^{-1} + A^T \sigma_w^{-1} A)^{-1}(\sigma_x^{-1} \bar{x} + A^T \sigma_w^{-1}(y - b)). \tag{5}
\]

3.2. Determination of HMM state sequence

The optimum state sequence \( q \) is determined by maximizing the output probability of a given speech acoustics. The output probability, given by eq.(1), is approximated by replacing the summation over all possible state sequences \( q \) by the maximum among all \( q \),

\[
P(y|\lambda) = \max_q \int P(y|x, q, \lambda)P(x|q, \lambda)P(q|\lambda)dx. \tag{5}
\]

Defining \( \sigma = (\sigma_x^{-1} + A^T \sigma_w^{-1} A)^{-1} \), we can write eq.(4) as

\[
\hat{x} = \bar{x} + \sigma A^T \sigma_w^{-1}(y - A\bar{x} - b). \tag{6}
\]

By substituting eq.(2), (3) and (6) into eq.(5), the output probability is given by

\[
P(y|\lambda) = \max_q P(q|\lambda) \frac{1}{(2\pi)^{N/2} |\sigma_y|^{1/2}} \times \exp \left[ -\frac{1}{2}(y - \bar{y})^T \sigma_y^{-1}(y - \bar{y}) \right] \times \int \frac{1}{(2\pi)^{M/2} |\sigma_x|^{1/2}} \times \exp \left[ -\frac{1}{2}(x - \bar{x})^T \sigma_x^{-1}(x - \bar{x}) \right] dx,
\]

where

\[
\bar{y} = A\bar{x} + b,
\]

\[
\sigma_y = A\sigma_x A^T + \sigma_w.
\]

As the integral part is equal to 1, then

\[
P(y|\lambda) = \max_q P(y|q, \lambda)P(q|\lambda).
\]

As the output probability has a form similar to the conventional HMM, the optimum state sequence \( q \) is determined from acoustic parameter vector \( y \) using the Viterbi algorithm.

4. SMOOTHING ARTICULATORY PARAMETERS

We propose a method of smoothing articulatory parameters that uses dynamic features [9]. In taking the dynamic features into account, we assume that articulatory parameter vector consists of articulatory parameters (static) and their velocity and acceleration (dynamic) and acoustic parameter vector consists of acoustic parameters and their velocity. Transform matrix \( R \) from static parameters to dynamic parameters is defined by

\[
R = \begin{bmatrix}
0_{m \times m} & 0_{m \times m} & I_{m \times m} \\
0_{m \times m} & -g_1 I_{m \times m} & g_1 I_{m \times m} \\
g_2 I_{m \times m} & -2g_2 I_{m \times m} & g_2 I_{m \times m}
\end{bmatrix},
\]
where $g_1$ and $g_2$ are the weight parameter, $0$ and $I$ are zero and identity matrix, and $m$ is the dimension of the static parameters of $x$. By maximizing a posteriori probability of articulatory parameter vector with dynamic features, the smoothed estimate of articulatory parameters $\hat{x}^*$ is given as

$$\hat{x}^* = (R^T \sigma_x^{-1} R + R^T A^T \sigma_w^{-1} A R)^{-1} \times (R^T \sigma_x^{-1} x + R^T A^T \sigma_w^{-1} (y - b)) .$$

5. TRAINING OF HMM PARAMETERS

Using simultaneously observed speech acoustics and articulatory movements, HMMs $\lambda$ is trained by maximizing the joint occurrence probability $P(x, y|\lambda)$. For models $\lambda$ and $\tilde{\lambda}$, the Q-function can be defined as

$$Q(\lambda, \tilde{\lambda}) = \sum_q P(x, y, q|\lambda) \log P(x, y, q|\tilde{\lambda}) .$$

Denoting a posteriori probability of being in state $j$ at time $t$ as $\gamma_t(j)$ and the probability of being in state $i$ at time $t - 1$ and in state $j$ at time $t$ as $\xi_t(i, j)$, we obtain the re-estimation equation for the articulatory-to-acoustic mapping as

$$H_j \sum_t \gamma_t(j)\mu_t\mu_t^T = \sum_t \gamma_t(j)y_t\mu_t^T .$$

Here, $H_j = [h_j A_j]$ and $\mu_t^T = [1|x_t^T]^T$. Similarly, the re-estimation equations for the mean $\pi_j$ and covariance $\sigma_{x_j}$, of $x$, the covariance $\sigma_{w_j}$ of $w$ in state $j$, and the probability $a_{ij}$ of the transition from state $i$ to state $j$ are given as

$$\pi_j = \frac{\sum_t \gamma_t(j)x_t}{\sum_t \gamma_t(j)} ,$$

$$\sigma_{x_j} = \frac{\sum_t \gamma_t(j)(x_t - \pi_j)(x_t - \pi_j)^T}{\sum_t \gamma_t(j)} ,$$

$$\sigma_{w_j} = \frac{\sum_t \gamma_t(j)(y_t - H_j\mu_t)(y_t - H_j\mu_t)^T}{\sum_t \gamma_t(j)} ,$$

$$a_{ij} = \frac{\sum_t \xi_{t-1}(i, j)}{\sum_t \gamma_{t-1}(i)} .$$

6. EXPERIMENT

We conducted experiments to evaluate the proposed method. In the experiments, articulatory movements and speech acoustics data were obtained from simultaneous observations constructed using an electro-magnetic articulograph (EMA) and acoustic recording of continuous speech utterances.

The articulatory data was collected using the EMA at a sampling rate of 250 Hz. The articulatory parameters are represented by the vertical and horizontal positions of eight points on the lower jaw, the upper and lower lips, the tongue (four positions), and the velum. Speech signal was sampled at 8 kHz and 24 mel-cepstral coefficients without the 0-th coefficient were obtained as acoustic parameters using a 32 ms Blackman window with a 4 ms frame.

Articulatory-acoustic recordings were made for 358 sentences (239,077 frames) spoken by a Japanese male. For these, 342 randomly selected ones were used as training data and the remaining 16 were used as test data for the evaluation. The type of HMM was 3-state left to right diphone models with no skips. The weight parameter $g_1$, $g_2$ was set at 0.5, 0.25, respectively.

We tested in two experimental conditions; namely, with and without phonemic information. In the former, the phoneme-specific HMM was temporarily assigned according to a given phoneme sequence. In the latter, the state sequence was determined for all possible sequences of the phoneme-specific HMM and a silent model. For both conditions, we evaluated the rms error between observed and estimated movements.

7. RESULT

Fig.2 shows an example of the observed and estimated movements of the vertical positions of the articulatory movements with/without phonemic information. The estimated articulatory movements match the observed movements well.

Fig.3 shows the rms error between the observed and estimated movements for each type of phoneme. In this figure, the error in 'Total' was the average rms error in an utterance. The error in 'Vowel', 'Labial', 'Alveolar' and 'Velar' was evaluated for all positions, the lip, the tongue tip, and the tongue blade and the velum, respectively. The error in 'Total' was 1.79 mm with phonemic information and 2.16 mm without phonemic information. The error in each type of phoneme was relatively small for labial, but large for velar.

8. CONCLUSION

We have presented an HMM-based speech production model for determining articulatory movements from speech acoustics. This model, which consists of an articulatory HMM for each phoneme and an articulatory-to-acoustic
Fig. 2. Observed (thin lines) and estimated (thick lines) articulatory movements of vertical positions of the lower jaw (LJ), the upper lip (UL), the lower lip (LL), four points on the tongue (T1-T4) from tip to blade and the velum (V). Left: with phonemic information. Right: without phonemic information.

Fig. 3. Average rms error for the each type of phoneme.

mapping for each HMM state, is efficient for implementing acoustic-to-articulatory inverse mapping. The average rms error is 1.79 mm with phonemic information and 2.16 mm without phonemic information.

9. REFERENCES


