## Integration of Acoustic Modeling and Mel-cepstral analysis for HMM-based Speech Synthesis

Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda

### Journal or Publication Title

### Page Range
7883-7887

### Year
2013

### URL
http://id.nii.ac.jp/1476/00004649/
doi: 10.1109/ICASSP.2013.6639199(http://dx.doi.org/10.1109/ICASSP.2013.6639199)
INTEGRATION OF ACOUSTIC MODELING AND MEL-CEPSRAL ANALYSIS
FOR HMM-BASED SPEECH SYNTHESIS

Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, Japan

ABSTRACT

In this paper, a novel approach for integrating acoustic modeling and mel-cepstral analysis is proposed. The aim of HMM-based speech synthesis is to model speech waveforms with a statistical model. However, the conventional techniques divide the modeling process into two steps: the frame by frame feature extraction step and the acoustic modeling step. Although it is reasonably effective, the deterioration of speech quality is caused by the divide of the objective function. In this paper, we propose an approach to modeling them as an integrative model and show the possibility of improving synthesized speech.

Index Terms— integrative model, HMM-based speech synthesis, acoustic modeling, mel-cepstral analysis, trajectory HMM

1. INTRODUCTION

HMM-based speech synthesis was proposed to enable machines to speak naturally like humans [1]. In this method, spectral and F0 features are extracted and modeled with a statistical technique. Recent large systems are often constructed with combinations of some modules that use statistical models. A famous example is language and acoustic models for speech recognition systems. Recently, the integration of these statistical models is an important research subject. In text-to-speech (TTS) systems, an approach integrating text analysis and speech synthesis modules was proposed [2]. It can optimize linguistic and acoustic models simultaneously. As the essential aim of TTS is to synthesize speech from given texts, the integration of these statistical models is a desirable future of TTS systems. In HMM-based TTS training, spectral envelope, fundamental frequency, and duration are modeled simultaneously by using the corresponding HMMs. Mel-cepstral coefficients are widely used as statistical model parameters and estimated from a given input signal, \( x \), in the maximum likelihood (ML) sense:

\[
\hat{c} = \arg \max_{c} P(x|c) \tag{1}
\]

Extracted mel-cepstral coefficients are used for training HMMs with dynamic (“delta” and “delta-delta”) feature constraints. A speech waveform is finally synthesized from the generated spectral and excitation parameters via the source-filter based production model. Recently, the trajectory HMM was derived by reformulating the HMM. The model parameters \( \Delta \) are trained in the ML sense by using the static features:

\[
\hat{\Delta} = \arg \max_{\Delta} P(e|\Delta) \tag{2}
\]

\[\Delta \]

2.1. MEL-CEPTRAL ANALYSIS

The synthesis filter \( H(z) \) is represented by mel-cepstral coefficients \( c = [c(0), \cdots, c(M)]^T \) defined as frequency-transformed cepstral coefficients:

\[1\text{In section 2.1, } x \text{ and } c \text{ correspond to not an utterance but a frame. The frame index } t \text{ is abbreviated.}\]
where \( \alpha \) is a frequency warping parameter. If \( \alpha = 0 \), mel-cepstral coefficients are equivalent to cepstral coefficients.

For a given input signal, \( x = [x(0), \ldots, x(N-1)]^T \), the mel-cepstral coefficients are determined by minimizing a spectral evaluation function with respect to \( c \) [7],

\[
E(x, c) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \exp R(\omega) - R(\omega) - 1 \right\} d\omega
\]

where

\[
R(\omega) = \log I_N(\omega) - \log \left| \frac{H(e^{j\omega})}{H(e^{j\omega})^2} \right|^2
\]

and \( I_N(\omega) \) is the modified periodogram of weakly stationary process \( x(n) \) with a time window \( w(n) \) of length \( N \):

\[
I_N(\omega) = \frac{\sum_{n=0}^{N-1} w(n)x(n)e^{-j\omega n}}{\sum_{n=0}^{N-1} w^2(n)}
\]

Mel-cepstral coefficients are determined easily by using an iterative algorithm (e.g., the Newton-Raphson method) because \( E(x, c) \) is convex with respect to \( c \).

When \( x(n) \) is assumed to be a zero-mean Gaussian process, the likelihood can be approximated by

\[
P(x|c) \approx \exp \left[ -\frac{N}{2} \log(2\pi) - \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \log \left| \frac{H(e^{j\omega})}{H(e^{j\omega})^2} \right|^2 \right\} d\omega \right]
\]

and, accordingly, minimization of \( E(x, c) \) corresponds to the maximization of \( P(x|c) \).

### 2.2. TRAJECTORY HMM

Let a spectral feature vector sequence be \( o = [o_1^T, \ldots, o_T^T]^T \), where \( o_t = [c_t^T, \Delta c_t^T, \Delta^2 c_t^T]^T \) includes not only static but also dynamic features. Mel-cepstral coefficients \( c_t \) is a \( M+1 \) dimensional vector, and \( T \) is the number of frames. In the conventional model, the probability density of \( o \) is shown as \( P(o|q, \Lambda) \), where \( q = [q_1, q_2, \ldots, q_T] \) is a state sequence. However, the conventional model is mathematically improper in the sense of statistical modeling. In this model, the static and dynamic features are modeled as independent statistical variables. When it is used as a generative model, it allows inconsistent static and dynamic features. By imposing an explicit relationship between static and dynamic features, which is given by \( o = Wc \), where \( W \) is a \( 3(M+1)T \times (M+1)T \) window matrix as shown in Fig. 1, the conventional HMM is reformed as the trajectory HMM as:

\[
P(c|\Lambda) = \sum_{q} P(c|q, \Lambda) P(q|\Lambda)
\]

\[
P(c|q, \Lambda) = N(c|\bar{c}_q, \Sigma_q) = \frac{1}{Z} P(o|q, \Lambda)
\]

\[
P(q|\Lambda) = P(q_1|\Lambda) \prod_{t=2} P(q_t|q_{t-1}, \Lambda)
\]

where \( Z \) is a normalization term. In Eq. (10), \( \bar{c}_q \) and \( \Sigma_q \) are the \((M+1)T \times 1\) mean vector and the \((M+1)T \times (M+1)T\) temporal utterance covariance matrix given by \( q \), respectively. They are given by

\[
R_q \bar{c}_q = r_q
\]

\[
R_q = W^T \Sigma_q^{-1} W = P_q^{-1}
\]

\[
\mu_q = \left( \mu_{q_1}^T, \ldots, \mu_{q_T}^T \right)^T
\]

\[
\Sigma_q = \text{diag} \left[ \Sigma_{q_1}, \ldots, \Sigma_{q_T} \right]^T
\]

where \( N \) is the total number of state output PDFs, and \( \mu_i \) and \( \Sigma_i \) are the \((M+1)T \times 1\) mean vector and the \((M+1)T \times (M+1)T\) covariance matrix associated with the \( i \)-th state, respectively. The elements of \( W \) are given as regression window coefficients to calculate delta and delta-delta features as:

\[
\Delta^d c_t = \sum_{\tau=-L^d}^{\tau=L^d} w^{(d)}(\tau)c_{t+\tau}, \quad d = 1, 2
\]

\[
W = \begin{bmatrix} W_1 & W_2 & \ldots & W_T \end{bmatrix} \otimes I_{(M+1) \times (M+1)}
\]

\[
w_t^{(d)} = \begin{bmatrix} 0, \ldots, 0, w^{(d)}(-L^d_1), \ldots, w^{(d)}(0), \ldots, w^{(d)}(L^d_1), \ldots, 0 \end{bmatrix}^T
\]

where \( L^{(0)} = 0, w^{(0)} = 1 \), and \( \otimes \) denotes the Kronecker product for matrices.

Note that \( c \) is modeled by Gaussian distributions whose dimensionality is \((M+1)T\), and the covariance matrices \( P_q \) are generally full. As a result, the trajectory HMM can overcome the deficiencies of the HMM. It is also noted that the parameterization of
the trajectory HMM is completely the same as that of the HMM with the same model topology.

3. INTEGRATION ALGORITHM OF ACOUSTIC MODELING AND MEL-CEPSTRAL ANALYSIS

Figure 2 shows the difference between the conventional and the proposed approaches. In the proposed approach, we integrate the statistical mel-cepstral model \( P(x|c) \) and the statistical acoustic model \( P(c|\Lambda) \) as:

\[
\Lambda = \arg\max_{\Lambda} P(x|\Lambda) = \arg\max_{\Lambda} \int P(x, c|\Lambda) \, dc = \arg\max_{\Lambda} \int P(x|c) P(c|\Lambda) \, dc \tag{23}
\]

In the proposed method, the mel-cepstral coefficients are represented as a random variable, and the dynamic feature vectors cannot be calculated from static features. Therefore, a modeling technique including the extraction of dynamic features, like trajectory HMM, should be introduced. As far as the conventional method, the standard mel-cepstral analysis can be assumed as extracting 1-best mel-cepstral coefficients.

To train the proposed model, a lower bound of log marginal likelihood \( \mathcal{F} \) is maximized instead of the true likelihood. The lower bound \( \mathcal{F} \) is defined by using Jensen’s inequality:

\[
\mathcal{L}(x|\Lambda) = \log P(x|\Lambda) = \log \sum_{q} \int P(x|c) P(c,q|\Lambda) \, dc = \log \sum_{q} \int Q(c) \frac{P(x|c) P(c,q|\Lambda)}{Q(c,q)} \, dc = \log \sum_{q} \int Q(c) Q(q) \log \frac{P(x|c) P(c,q|\Lambda)}{Q(c) Q(q)} \, dc \geq \sum_{q} \int Q(c) Q(q) \log \frac{P(x|c) P(c,q|\Lambda)}{Q(c) Q(q)} \, dc = \mathcal{F} \tag{24}
\]

To overcome the difficulty of optimization, we assume that \( c \) and \( q \) are independent and that posterior distribution \( Q(c) \) and \( Q(q) \) approximate the true posterior distributions. The optimal posterior distributions can be obtained by maximizing the original objective function \( \mathcal{F} \) with the variational method as:

\[
Q(c) = \frac{1}{Z_c} P(x|c) \exp \sum_{q} Q(q) \log P(c,q|\Lambda) \tag{25}
\]

\[
Q(q) = \frac{1}{Z_q} P(q|\Lambda) \exp \int Q(c) \log P(c,q|\Lambda) \, dc \tag{26}
\]

where \( Z_c \) and \( Z_q \) are the normalization terms of \( Q(c) \) and \( Q(q) \), respectively. These optimizations can be effectively performed by iterative calculations as the Expectation and Maximization (EM) algorithm, which increases the value of objective function \( \mathcal{F} \) at each iteration until convergence.

3.1. Posterior Probabilities of Mel-cepstral coefficients

It is difficult to integrate \( Q(c) \) with respect to \( c \), so we approximate \( Q(c) \) as a Gaussian probability distribution. By using a Laplace approximation, \( Q(c) \) is represented as:

\[
Q(c) \simeq N(\hat{c}, A^{-1}) \tag{27}
\]

\[
\hat{c} = \arg\max_{c} Q(c) \quad \text{and} \quad A = \frac{N}{2} H \mid_{c=\hat{c}} + \sum_{q} Q(q) P_q^{-1} \tag{29}
\]

where

\[
H = -\frac{2}{N} \frac{\partial^2}{\partial c \partial c^T} \log P(x|c) \tag{30}
\]

\[
H_t = \frac{\partial^2}{\partial c_t \partial c_t^T} E(x_t, c_t) = -\frac{2}{N} \frac{\partial^2}{\partial c_t \partial c_t^T} \log P(x_t|c_t) \tag{32}
\]

In the standard mel-cepstral analysis, mel-cepstral coefficients for each frame can be estimated frame by frame independently. However, in the proposed method, mel-cepstral coefficients for all utterances should be estimated simultaneously. Thus, a large computation cost is required for this Newton-Raphson method.

3.2. Posterior Probabilities of State Sequences

The expectation with respect to \( c \) in Eq. (26) is given by

\[
\int Q(c) \log P(c,q|\Lambda) \, dc = \log \mathcal{N}(\hat{c}|c_q, P_q) - \frac{1}{2} \text{tr} \left( R_q A^{-1} \right) \tag{33}
\]

where the matrix \( R_q \) is a positive definite \((4L(M+1)+1)\)-diagonal band symmetric matrix. Thus, \( R_q \) can be decomposed into its Cholesky factorization:

\[
R_q = U_q^T U_q \tag{34}
\]

where \( U_q \) is an upper-triangular \((2L(M+1)+1)\)-diagonal matrix. Elements of \( U_q \) are calculated in a recursive manner and depend only on substrates from time 1 to \( t + 2L \).

\[
(R_q A^{-1})^{(t,i)} = \left( U_q^T U_q A^{-1} \right)^{(t,i)} = \left( U_q A^{-1} U_q^T \right)^{(t,i)} = \sum_{i=1}^{t+2L} \sum_{j=t}^{t+2L} U_q^{(i,j)} \left( A^{-1} \right)^{(i,j)} U_q^{(i,j)} \tag{35}
\]
Thus, the delayed decision Viterbi algorithm [4] can be applied to the proposed method.

3.3. Update Model Parameters

Model parameters $m$ and $\phi$ are defined by concatenating the mean vectors and covariance matrices of all unique Gaussian components in the model set as:

$$m = \left[ \mu_1^T, \mu_2^T, \cdots, \mu_N^T \right]^T$$  (36)

$$\phi = \left[ \Sigma_1^T, \Sigma_2^T, \cdots, \Sigma_N^T \right]^T$$  (37)

where $\mu_n$ and $\Sigma_n$ are the mean vector and covariance matrix of the $n$-th unique Gaussian component in the model set, and $N$ is the total number of Gaussian components in the model set, respectively.

By setting the partial derivative of $F$ with respect to $m$ to 0, a set of linear equations for determining $m$ maximizing $F$ are obtained as:

$$\sum_{\forall q} Q(q) S_q^T W \Phi_q W^T S_q^{-1} m = \sum_{\forall q} Q(q) S_q^T W \bar{c}$$  (38)

where

$$\mu_q = S_q m$$  (39)

$$\Phi_q^{-1} = \text{diag}(\phi)$$  (40)

$$\Sigma_q^{-1} = \text{diag}(S_q \phi)$$  (41)

$$S_q \Phi_q^{-1} = \Sigma_q^{-1} S_q$$  (42)

In the above equations, $S_q$ is a $3(M+1) T \times 3(M+1) T$ matrix whose elements are 0 or 1 determined by the Gaussian component sequence $q$.

For maximizing $F$ with respect to $\phi$, a gradient method is applied by using its partial derivative

$$\frac{\partial F}{\partial \phi} = \sum_{\forall q} Q(q) \left[ \frac{1}{2} S_q^T \text{diag}^{-1}\left\{ W P_q W^T - W A^{-1} W^T \right\} - W \bar{c} \bar{c}^T W^T + 2 \mu_q \bar{c}^T W^T + \mu_q \bar{c}^T W^T \right]$$

(43)

because Eq. (43) is not a quadratic function of $\phi$.

5. CONCLUSION

In this paper, we defined a novel kind of acoustic model for modeling speech waveforms directly by integrating the mel-cepstral analysis and the acoustic modeling. In experiments, the objective and subjective evaluation scores of proposed models were equivalent to or higher than that of the trajectory HMMs. These results indicate the possibility of improving the quality of synthesized speech. Experiments on larger data sets will be future work.

6. ACKNOWLEDGMENTS

The research leading to these results was partly funded by the Core Research for Evolutionary Science and Technology (CREST) from the Japan Science and Technology Agency (JST).
7. REFERENCES


