

Mixture Density Models Based on Mel-Cepstral Representation of Gaussian Process

Toru TAKAHASHI^{†a)}, Keiichi TOKUDA^{††}, Takao KOBAYASHI[†],
and Tadashi KITAMURA[†], *Regular Members*

SUMMARY This paper defines a new kind of a mixture density model for modeling a quasi-stationary Gaussian process based on mel-cepstral representation. The conventional AR mixture density model can be applied to modeling a quasi-stationary Gaussian AR process. However, it cannot model spectral zeros. In contrast, the proposed model is based on a frequency-warped exponential (EX) model. Accordingly, it can represent spectral poles and zeros with equal weights, and, furthermore, the model spectrum has a high resolution at low frequencies. The parameter estimation algorithm for the proposed model was also derived based on an EM algorithm. Experimental results show that the proposed model has better performance than the AR mixture density model for modeling a frequency-warped EX process.

key words: *Gaussian mixture model, statistical framework, mel-cepstrum, EM algorithm*

1. Introduction

In applications such as signal enhancements, classifications, and verifications, a signal is often modeled in a statistical framework. An autoregressive (AR) spectral analysis algorithm, in which a signal is assumed to be an AR Gaussian process, has been used extensively to model a stationary process. When a signal can be assumed to be quasi-stationary, i.e., it is non-stationary whereas the short-term segments of a signal can be regarded as stationary processes, the AR mixture density models and the autoregressive hidden Markov model (AR-HMM) [1] can be applied. However, they cannot model spectral zeros; in some cases, spectral zeros are important in characterizing a signal, i.e., nasal vowels and nasal consonants. Although the autoregressive moving average (ARMA) models can determine poles and zeros simultaneously, they are not successful for a variety of reasons, e.g., stability of obtained system function and convergence of an iterative algorithm.

This paper defines a new kind of a mixture density model for modeling a quasi-stationary Gaussian process based on mel-cepstral representation of the power

spectrum; the paper also derives an EM algorithm for the model parameter estimation. The proposed model is based on the exponential (EX) model that can represent spectral poles and zeros with equal weights. Furthermore, it has a high resolution at low (or high) frequencies with an appropriate choice of a frequency-warping parameter since it uses a frequency-warped EX model.

In order to model a quasi-stationary process, the Gaussian mixture model (GMM) is often applied to spectral parameter vectors (e.g., cepstral coefficient vectors) that are extracted from the short-term segments of the signal to be modeled. In this approach, however, two different criteria must be used; one for extracting spectral parameters and another for modeling a mixture density. The proposed technique can be viewed as a unified approach to the problem; in the proposed approach, only one criterion, likelihood, is used for the estimation of the spectral model parameters from the observed signal.

The standard hidden Markov model assumes that each state generates observations according to a Gaussian or Gaussian mixture model. On the other hand, some HMMs with extended observation densities are summarized in [2]. The AR observation model embedded in an HMM as described in [2] is the AR-HMM. Similarly, the frequency-warped EX model embedded in an HMM [3] is the EX-HMM. It should be noted that the proposed model is equivalent to the EX-model when it is a single-state multi-mixture model.

This paper is organized as follows. Section 2 summarizes the mel-cepstral analysis technique [4], which can be viewed as a spectral estimation technique for frequency-warped EX Gaussian processes. Section 3 defines the frequency-warped EX mixture density model and derives an EM algorithm for model parameter estimation. Finally, experimental results and conclusions are given in Sects. 4 and 5, respectively.

2. Mel-Cepstral Analysis Technique [4]

The synthesis filter $H(z)$ is represented by mel-cepstral coefficients $c(m)$, ($m = 0, 1, \dots, M$) defined as frequency-transformed cepstral coefficients:

Manuscript received November 28, 2002.

Manuscript revised March 4, 2003.

Final manuscript received April 11, 2003.

[†]The authors are with the Department of Electrical and Electronic Engineering, Nagoya Institute of Technology, Nagoya-shi, 466-8555 Japan.

^{††}The author is with the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226-8502 Japan.

a) E-mail: tall@ics.nitech.ac.jp

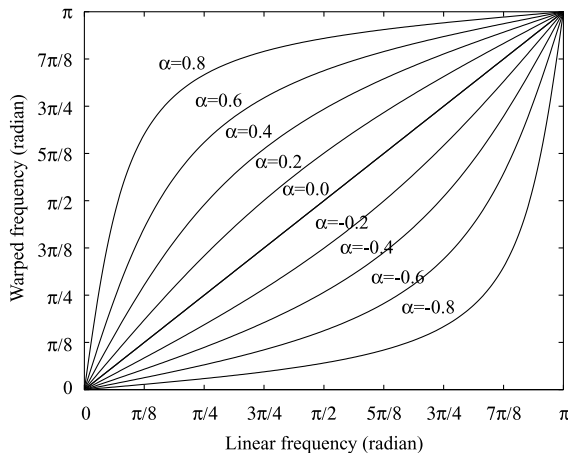


Fig. 1 Frequency warping.

$$H(z) = \exp \sum_{m=0}^M c(m) \tilde{z}^{-m}, \quad (1)$$

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad (2)$$

where

$$|\alpha| < 1. \quad (3)$$

If $\alpha = 0$, mel-cepstral coefficients are equivalent to cepstral coefficients. Figure 1 shows the frequency warping function for different values of α . The vertical axis gives warped frequencies and the horizontal axis gives the linear frequencies. If $\alpha > 0$, system function defined as Eq. (1) has a high resolution at low frequencies, and if $\alpha < 0$, it has a high resolution at high frequencies.

For given input signal, $\mathbf{x} = [x(0), \dots, x(N-1)]^T$, the mel-cepstral coefficients, $\mathbf{c} = [c(0), \dots, c(M)]^T$, are determined by minimizing a spectral evaluation function [5]

$$E(\mathbf{x}, \mathbf{c}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{R(\omega) - \log R(\omega) - 1\} d\omega, \quad (4)$$

with respect to \mathbf{c} , where

$$R(\omega) = \frac{I(\omega)}{|H(e^{j\omega})|^2}, \quad (5)$$

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

$$I(\omega) = \frac{1}{W} \left| \sum_{n=0}^{N-1} w(n)x(n)e^{-j\omega n} \right|^2, \quad (6)$$

$$W = \sum_{n=0}^{N-1} w^2(n). \quad (7)$$

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

When $x(n)$ is assumed to be a zero-mean Gaussian process, the likelihood of \mathbf{c} for \mathbf{x} , $P(\mathbf{x}|\mathbf{c})$, can be approximated by

$$P(\mathbf{x}|\mathbf{c}) \cong \exp \left[-\frac{N}{2} \left[\log(2\pi) + \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \log |H(e^{j\omega})|^2 + \frac{I(\omega)}{|H(e^{j\omega})|^2} \right\} d\omega \right] \right], \quad (8)$$

and accordingly minimization of $E(\mathbf{x}, \mathbf{c})$ corresponds to maximization of $P(\mathbf{x}|\mathbf{c})$. It should be noted that the spectral evaluation function of mel-cepstral analysis has the same form as that of LPC analysis [6]. Furthermore, taking the gain factor G outside from $H(e^{j\omega})$ indicates that the minimization of $E(\mathbf{x}, \mathbf{c})$ with respect to \mathbf{c} is equivalent to both minimization of residual energy and maximization of the prediction gain.

3. Mixture Density Model Based on Mel-Cepstral Representation

For a spectral parameter $\mathbf{a} = [a(0), a(1), \dots, a(P)]^T$, the probability density function (pdf) of the conventional GMM λ is represented by the sum of weighted Gaussian pdfs and is defined as

$$\begin{aligned} P(\mathbf{a}|\lambda) &= \sum_{k=1}^K w_k \cdot P(\mathbf{a}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \sum_{k=1}^K w_k \cdot \frac{\exp \left\{ -\frac{1}{2} (\mathbf{a} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{a} - \boldsymbol{\mu}_k) \right\}}{\sqrt{(2\pi)^{P+1} |\boldsymbol{\Sigma}_k|}}, \\ &= \sum_{k=1}^K w_k \cdot \exp \left[-\frac{P+1}{2} \left\{ \log(2\pi) + \frac{1}{P+1} \left(\log |\boldsymbol{\Sigma}_k| + (\mathbf{a} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{a} - \boldsymbol{\mu}_k) \right) \right\} \right] \end{aligned} \quad (9)$$

where

$$\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_K\}, \quad (10)$$

$$\lambda_k = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, \quad (k = 1, 2, \dots, K). \quad (11)$$

The parameter set of the k -th component is represented by $\lambda_k = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. The weight of the k -th Gaussian component, the mean vector of the k -th Gaussian component, and the covariance matrix of the k -th Gaussian component are represented by w_k , $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$, respectively. By substituting the pdf of the frequency warped EX model for $P(\mathbf{a}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, a new mixture density model can be derived.

3.1 Gaussian Mixture Model Based on Mel-Cepstral Representation

The pdf of the proposed mixture model is represented

by the sum of weighted pdfs defined by (8). Thus, the pdf of the proposed model λ is defined as

$$\begin{aligned} P(\mathbf{x}|\lambda) &= \sum_{k=1}^K w_k \cdot P(\mathbf{x}|\mathbf{c}_k) \\ &= \sum_{k=1}^K w_k \cdot \exp \left[-\frac{N}{2} \left\{ \log(2\pi) \right. \right. \\ &\quad \left. \left. + \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\log |H_k(e^{j\omega})|^2 + \frac{I(\omega)}{|H_k(e^{j\omega})|^2} \right) d\omega \right\} \right], \quad (12) \end{aligned}$$

where

$$\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_K\}, \quad (13)$$

$$\lambda_k = \{w_k, \mathbf{c}_k\}, \quad (k = 1, 2, \dots, K). \quad (14)$$

The parameter set of the k -th component is represented by $\lambda_k = \{w_k, \mathbf{c}_k\}$. The weight of the k -th component is represented by w_k . The mel-cepstral coefficient vector $\mathbf{c}_k = [c_k(0), c_k(1), \dots, c_k(M)]^T$ for the k -th component represents the system function $H_k(z)$ as follows:

$$H_k(z) = \exp \sum_{m=0}^M c_k(m) \tilde{z}^{-m}. \quad (15)$$

It should be noted that although the pdf of the proposed model λ has a form similar to the AR mixture density model, it can represent spectral poles and zeros with equal weights and has a high resolution at low or high frequencies with an appropriate choice of the frequency warping parameter α . Furthermore, this model has fewer model parameters than the conventional model for the same mixture number and the same number of spectral parameters. In case of K mixtures and M -th order analysis, the proposed model has $K \cdot (1 + (M + 1))$ model parameters, and conventional GMM with a diagonal covariance matrix and conventional GMM with a full covariance matrix have $K \cdot (1 + 2 \cdot (M + 1))$ model parameters and $K \cdot (1 + (M + 1) + \frac{1}{2}(M + 1) \cdot (M + 2))$ model parameters, respectively.

3.2 Algorithm for Model Parameter Estimation

This section describes a model parameter estimation algorithm. The model parameter set λ^* is given by maximizing the log likelihood for training vector set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$, that is

$$\lambda^* = \arg \max_{\lambda} \log P(\mathbf{X}|\lambda) \quad (16)$$

$$= \arg \max_{\lambda} \sum_{t=1}^T \log P(\mathbf{x}_t|\lambda). \quad (17)$$

It should be noted that \mathbf{x}_t is the t -th signal segment, which can be regarded as a stationary process. However, Eq. (17) is difficult to solve analytically. Therefore, the model parameter is estimated based on an

Step 0. Begin with an initial model, $\lambda(0)$, and set $0 \rightarrow n$.
Step 1. Calculate the log likelihood, L_n , of the given model, $\lambda(n)$ for the training vector set \mathbf{X} .
Step 2. Reestimate the model parameter set, $\lambda(n+1)$, by maximizing $Q(\lambda(n), \lambda(n+1))$ with respect to $\lambda(n+1)$.
Step 3. If the log likelihood $L_{(n+1)}$ of the reestimated model $\lambda(n+1)$ has changed by a small enough amount since the last iteration, stop the algorithm. Otherwise set $n+1 \rightarrow n$ and go to step 1.

Fig. 2 The proposed model parameter estimation algorithm.

EM algorithm. The log likelihood of the reestimated parameter set λ' , which is given by maximizing the Q-function, always increases. The sub-optimal model parameters can be estimated by maximizing the Q-function iteratively. A Q-function is defined as

$$Q(\lambda, \lambda') = \sum_{t=1}^T Q_t(\lambda, \lambda'), \quad (18)$$

$$Q_t(\lambda, \lambda') = \sum_{q_t \in \{1, 2, \dots, K\}} P(q_t | \mathbf{x}_t, \lambda) \log P(\mathbf{x}_t, q_t | \lambda'), \quad (19)$$

where

$$P(k | \mathbf{x}_t, \lambda) = \frac{w_k \cdot P(\mathbf{x}_t | \mathbf{c}_k)}{\sum_{i=1}^K w_i \cdot P(\mathbf{x}_t | \mathbf{c}_i)}, \quad (20)$$

$$P(\mathbf{x}_t, k | \lambda) = w_k \cdot P(\mathbf{x}_t | \mathbf{c}_k). \quad (21)$$

Figure 2 outlines the algorithm.

At Step 0, the initial model parameters are estimated by mel-cepstral vector quantization (MC-VQ) [7] in a manner similar to conventional GMM.

At Step 1, a log likelihood for training vector set, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$, is calculated using Eq. (12) as follows:

$$L_n = \log P(\mathbf{X}|\lambda) \quad (22)$$

$$= \sum_{t=1}^T \log P(\mathbf{x}_t|\lambda) \quad (23)$$

At Step 2, the parameter set $\lambda(n+1)$ is reestimated by maximizing the Q-function $Q(\lambda(n), \lambda(n+1))$ with respect to $\lambda(n+1)$. Q-function $Q(\lambda, \lambda')$ can be written as

$$Q(\lambda, \lambda') = Q^{(w)}(\lambda, \lambda') + Q^{(c)}(\lambda, \lambda'), \quad (24)$$

$$\begin{aligned} Q^{(w)}(\lambda, \lambda') &= \sum_{t=1}^T \sum_{q_t \in \{1, \dots, K\}} P(q_t | \mathbf{x}_t, \lambda) \log w'_{q_t}, \quad (25) \end{aligned}$$

$$Q^{(c)}(\lambda, \lambda') = \sum_{t=1}^T \sum_{q_t \in \{1, \dots, K\}} P(q_t | \mathbf{x}_t, \lambda) \log P(\mathbf{x}_t | \mathbf{c}'_{q_t}), \quad (26)$$

where $\mathbf{w} = [w_1, \dots, w_K]$ and $\mathbf{c} = \{c_1, \dots, c_K\}$. Equation (25) has the form of $\sum_{i=1}^N u_i \log y_i$, which attains its unique maximum point

$$y_i = \frac{u_i}{\sum_{j=1}^N u_j} \quad (27)$$

under the constraint $\sum_{i=1}^N y_i = 1$. Therefore, the parameter w_i which maximizes Eq. (25), subject to the stochastic constraints

$$\sum_{i=1}^N w'_i = 1 \quad (28)$$

can be derived as

$$w'_k = \frac{1}{T} \sum_{t=1}^T P(k | \mathbf{x}_t, \lambda). \quad (29)$$

The reestimation equation for \mathbf{c} is given by solving equation

$$\frac{\partial Q^{(c)}(\lambda, \lambda')}{\partial \mathbf{c}'_k} = \mathbf{0} \quad (k = 1, \dots, K) \quad (30)$$

Although Eq. (30) is difficult to solve analytically, the Q-function can be maximized by using iterative algorithm because the Q-function is convex with respect to \mathbf{c}_k . The maximization problem can be written as

$$\mathbf{c}'_k = \arg \max_{\mathbf{c}'_k} Q(\lambda, \lambda') \quad (31)$$

By omitting terms independent of \mathbf{c}'_k , it can be written as

$$\mathbf{c}'_k = \arg \max_{\mathbf{c}'_k} \sum_{t=1}^T [P(k | \mathbf{x}_t, \lambda) \cdot \log P(\mathbf{x}_t | \mathbf{c}'_k)]. \quad (32)$$

By substituting Eq. (8), it is written as

$$\begin{aligned} \mathbf{c}'_k &= \arg \max_{\mathbf{c}'_k} \sum_{t=1}^T \left[P(k | \mathbf{x}_t, \lambda) \cdot \left[-\frac{N}{2} \left\{ \log(2\pi) \right. \right. \right. \\ &\quad \left. \left. + \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\log |H'_k(e^{j\omega})|^2 + \frac{I_t(\omega)}{|H'_k(e^{j\omega})|^2} \right) d\omega \right\} \right] \right] \\ &= \arg \max_{\mathbf{c}'_k} \left[-\frac{N}{2} \left\{ \log(2\pi) \cdot \left\{ \sum_{t=1}^T P(k | \mathbf{x}_t, \lambda) \right\} \right. \right. \\ &\quad \left. \left. + \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\left\{ \sum_{t=1}^T P(k | \mathbf{x}_t, \lambda) \right\} \cdot \log |H'_k(e^{j\omega})|^2 \right. \right. \right. \\ &\quad \left. \left. + \frac{\sum_{t=1}^T \{P(k | \mathbf{x}_t, \lambda) \cdot I_t(\omega)\}}{|H'_k(e^{j\omega})|^2} \right) d\omega \right\} \right]. \quad (33) \end{aligned}$$

From Eq. (20),

$$\begin{aligned} \mathbf{c}'_k &= \arg \max_{\mathbf{c}'_k} \exp \left[-\frac{N}{2} \left\{ \log(2\pi) \right. \right. \\ &\quad \left. \left. + \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\log |H'_k(e^{j\omega})|^2 + \frac{\bar{I}_k(\omega)}{|H'_k(e^{j\omega})|^2} \right) d\omega \right\} \right], \quad (34) \end{aligned}$$

where $I_t(\omega)$ is the modified periodogram of an input speech segment \mathbf{x}_t with a time window $w(n)$ of length N :

$$I_t(\omega) = \frac{1}{W} \left| \sum_{n=0}^{N-1} w(n) x_t(n) e^{-j\omega n} \right|^2, \quad (35)$$

$$W = \sum_{n=0}^{N-1} w^2(n), \quad (36)$$

$$\bar{I}_k(\omega) = \frac{\sum_{t=1}^T P(k | \mathbf{x}_t, \lambda) I_t(\omega)}{\sum_{t=1}^T P(k | \mathbf{x}_t, \lambda)}. \quad (37)$$

If $\bar{I}_k(\omega)$ is regarded as the periodogram, solving the problem (34) is equivalent to maximizing Eq. (8). Therefore, the algorithm for the mel-cepstral analysis technique [4] can be used in order to solve the problem (34).

4. Experiments

In order to show the effectiveness of the proposed model, the spectra estimated by the proposed method are compared with ones estimated by the AR mixture density model (single state multi-mixture AR-HMM).

4.1 Estimation of Pseudo EX Gaussian Process

To generate the signal to be analyzed, the values of $\lambda = \{w_1, \dots, w_K, c_1, \dots, c_K\}$ were selected randomly and $M = 18$. The number of mixtures, K , was 32. The frequency-warping parameter α was set to 0.41. The signal was generated by exciting the MLSA filter with a white Gaussian noise for each \mathbf{c}_k (Fig. 3). The signals were switched according to the mixture probabilities

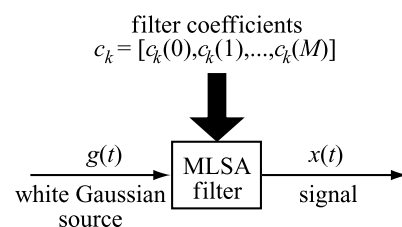


Fig. 3 Generating artificial signal.

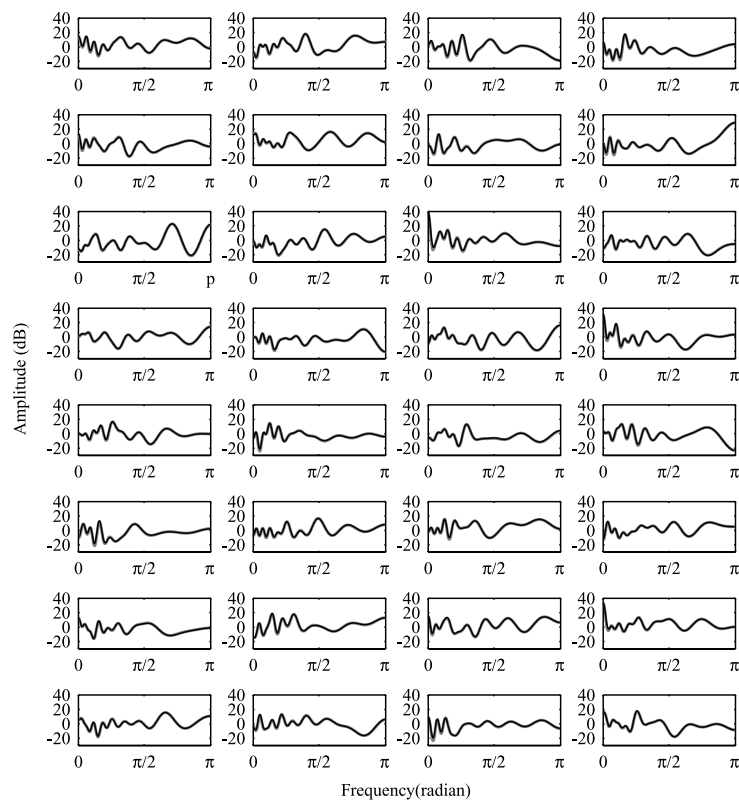


Fig. 4 Log amplitude responses estimated using the proposed model ($\alpha = 0.41$).

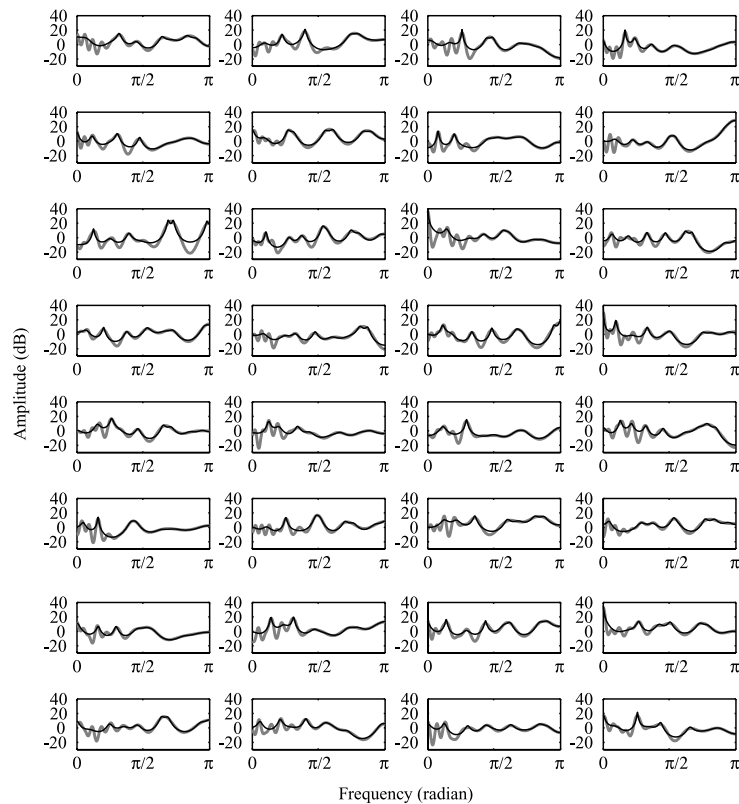


Fig. 5 Log amplitude responses estimated using AR mixture density model.

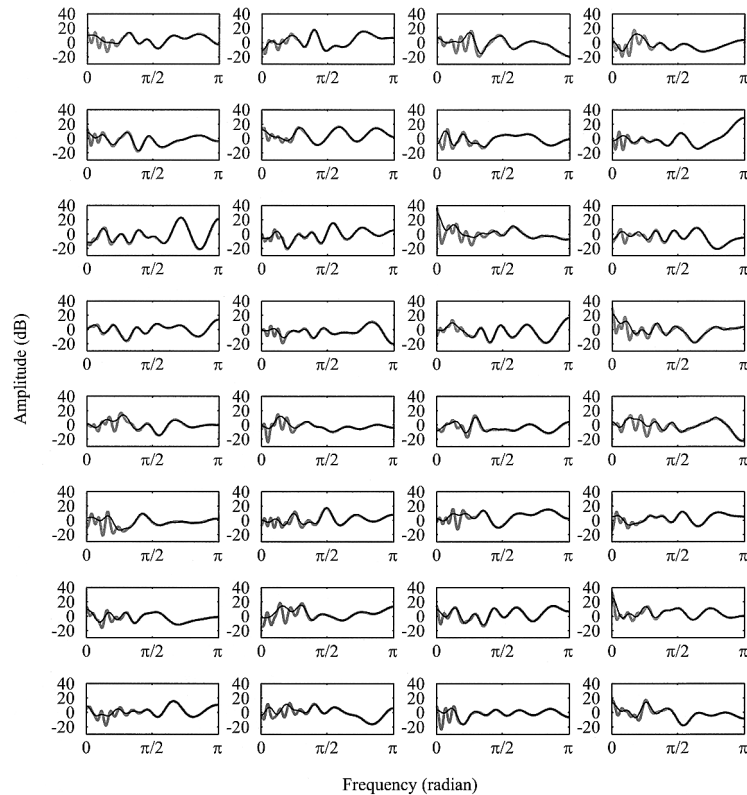


Fig. 6 Log amplitude responses estimated with the proposed model ($\alpha = 0$).

$\{w_1, \dots, w_K\}$. The segment length was assumed to be 320 sample points, and in this experiment, switching is assumed to occur at segment boundaries.

In Figs. 4 and 5, the vertical axis gives the log spectral amplitudes and the horizontal axis gives the frequencies. Figure 4 shows the spectra calculated from $\{c_1, \dots, c_K\}$ estimated by the model parameter estimation algorithm for the proposed model (black lines) with $K = 32$, $M = 18$, and $\alpha = 0.41$. The true spectra are also plotted with gray lines. The result for the AR mixture density model with $K = 32$ and $M = 18$ is shown in Fig. 5.

Figure 4 indicates that the unknown spectra were estimated accurately. On the other hand, Fig. 5 shows that the AR mixture density model cannot estimate spectral valleys, and especially at low frequencies.

Figure 6 shows the result for the proposed model with $K = 32$, $M = 18$, and $\alpha = 0$. In this case, it cannot estimate spectral valleys at low frequencies. From Figs. 4 and 6 indicate that, by choosing the value of α , the frequency resolution can be changed according to characteristics of the signal to be modeled.

4.2 Phonetic Model Training Experiments

A frequency-warped EX mixture density model is considered to be a suitable model for speech signals because a frequency-warped EX mixture density model

Table 1 Experimental conditions.

Database	ATR Japanese Speech Database C-set 216 words
Speakers	male 5 / female 5
Sampling rate	16 kHz
Frame length	32 ms
Frame period	5 ms
Order	$M = 18$
α	0.41

can define the speech signal model based on the mel-frequency scale, e.g., for the speech signals sampled at 8, 10, and 16 kHz, a frequency-warped EX mixture density model is defined in mel-scale frequency domain by setting $\alpha = 0.31, 0.35$, and 0.41 , respectively. Spectral poles and zeros in particular are known to play an important role in nasalized phonemes, i.e., /m/, /n/, and /N/. Therefore, the proposed model should be able to model those phonemes with greater likelihood than the AS model. In this section, several nasal phoneme models were trained based on the proposed model and the AR model. Experimental conditions are shown in Table 1. Figure 7 shows average log likelihood. The vertical axis gives the number of mixtures and the horizontal axis gives the average log likelihood. The results indicate that the log likelihood of proposed models for training signals is higher than that of AR models.

Figure 8 shows the log spectra calculated from mixture components c_k ($k = 1, 2, 3, 4$) for a phoneme

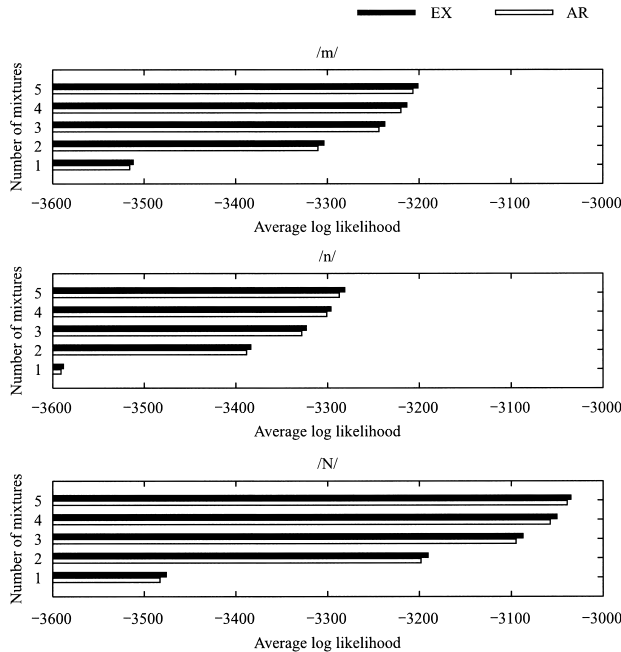


Fig. 7 Average log likelihood.

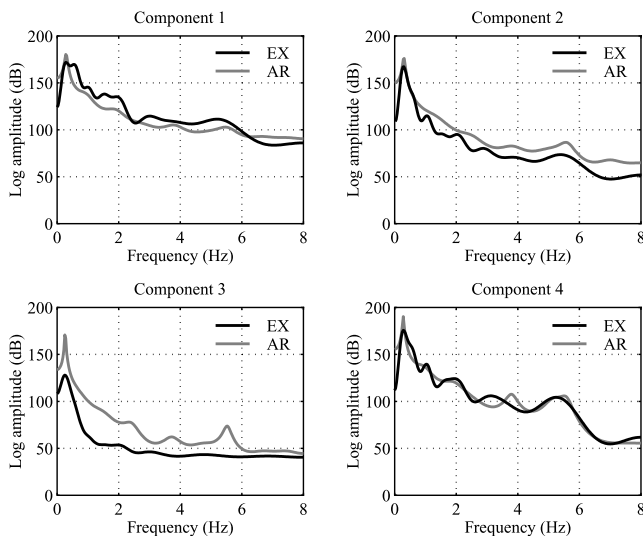


Fig. 8 Estimated log amplitude responses by 4 mixture models for phoneme /m/.

/m/ spoken by a female (f101). The vertical axis gives log amplitudes and the horizontal axis gives frequencies. The black lines and the gray lines represent the log spectra for the EX-GMM and the AR-GMM, respectively. The log spectra of the phoneme /m/ are known to have spectral valleys around 1 kHz [8]. The log spectra of the components 1, 2, and 4 for the EX-GMM display characteristics of the phoneme /m/ very well. However, those for the AR-GMM cannot represent spectral valleys around 1 kHz. This may be considered as the reason why the EX-GMM for the signals achieves greater likelihood than the AR-GMM in Fig. 7.

5. Conclusions

This paper has defined a new kind of mixture density model for modeling a quasi-stationary Gaussian process based on mel-cepstral representation of the power spectrum; the paper also derived a parameter estimation algorithm for the proposed model based on an EM algorithm. Results have indicated that the proposed model has better performance than an AR mixture density model for modeling a frequency-warped EX processes. Future works include applying the proposed model to practical applications, e.g., a signal enhancement, a classification, and a verification.

References

- [1] B.H. Juang and L.R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-33, no. 6, pp. 1404–1413, Dec. 1985.
- [2] W.D. Penny and S.J. Roberts, "Hidden Markov models with extended observation densities," Technical Report, Neural Systems Research Group, Imperial College of Science, Technology and Medicine, Oct. 1998.
- [3] T. Takahashi, K. Tokuda, T. Kobayashi, and T. Kitamura, "Training algorithm of HMMs based on mel-cepstral representation," *Proc. Autumn Meeting of ASJ*, vol. 1, pp. 5–6, Oct. 2001.
- [4] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. ICASSP*, vol. 1, pp. 137–140, March 1992.
- [5] S. Imai and C. Furuichi, "Unbiased estimator of log spectrum and its application to speech signal processing," *Proc. EURASIP*, pp. 203–206, Sept. 1988.
- [6] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *IECE Trans. Fundamentals (Japanese Edition)*, vol. J53-A, no. 1, pp. 35–42, Jan. 1970. Translation: R.W. Schafer and J.D. Markel, eds., *Speech Analysis*, pp. 295–302, IEEE Press, New York, 1979.
- [7] T. Takahashi, K. Tokuda, T. Kobayashi, and T. Kitamura, "Vector quantization of mel-cepstral coefficients based on a statistical measure," *Proc. ISAPCS*, vol. 2, pp. 692–695, Nov. 2000.
- [8] K.N. Stevens, *Acoustic Phonetics*, The MIT Press, 1998.



Toru Takahashi was born in Akita, Japan, in 1973. He received the B.E. degree in Computer Science, and the M.E. degree in the Department of Electrical and Electronic Engineering from the Nagoya Institute of Technology, Nagoya Japan, in 1996 and 1998, respectively. He is currently studying at the Department of Electrical and Electronic Engineering, Nagoya Institute of Technology. His research interests include digital signal processing, speech analysis and synthesis, and speech recognition. He is a member of ASJ and IPSJ.

processing, speech analysis and synthesis, and speech recognition. He is a member of ASJ and IPSJ.



Keiichi Tokuda was born in Nagoya, Japan, in 1960. He received the B.E. degree in Electrical and Electronic Engineering from the Nagoya Institute of Technology, Nagoya, Japan, and the M.E. and Dr.Eng. degrees in Information Processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a research Associate at the Department of Electronic and Electric Engineering,

Tokyo Institute of Technology. Since 1996 he has been with the Department of Computer Science, Nagoya Institute of Technology as Associate Professor. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. His research interests include speech coding, speech synthesis and recognition and multimodal signal processing. He is a member of IEEE, ISCA, IPSJ, ASJ and JSAI.



Takao Kobayashi was born in Niigata, Japan, in 1955. He received the B.E. degree in Electrical Engineering, the M.E. and Dr.Eng. degrees in Information Processing from Tokyo Institute of Technology, Tokyo, Japan, in 1977, 1979, and 1982, respectively. In 1982, he joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology as a Research Associate.

He became an Associate Professor at the same Laboratory in 1989. He is currently a Professor of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. His research interests include speech analysis and synthesis, speech coding, speech recognition, and multimodal interface. He is a member of IEEE, ISCA, IPSJ and ASJ.



Tadashi Kitamura was born in Ishikawa, Japan, in 1950. He received the B.E. degree in Electrical and Electronic Engineering from the Nagoya Institute of Technology, and the M.E. and Dr.Eng. degrees from the Tokyo Institute of Technology, in 1973, 1975, and 1978, respectively. He is currently a Professor at Interdisciplinary Graduate School of Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan. His research inter-

ests include digital signal processing, speech analysis and synthesis, image synthesis, and multimodal speech recognition. He is a member of IEEE, ISCA, IPSJ and ASJ.