COMPARISON OF AUTO-REGRESSIVE, NON-STATIONARY EXCITED SIGNAL PARAMETER ESTIMATION METHODS

Akira Sasou[1], Masataka Goto[1], Satoru Hayamizu[2], Kazuyo Tanaka[3]
[1]National Institute of Advanced Industrial Science and Technology (AIST)
{a-sasou,m.goto}@aist.go.jp
[2]Faculty of Engineering, Gifu University
hayamizu@info.gifu-u.ac.jp
[3]Institute of Library and Information Science, University of Tsukuba
ktanaka@ulis.ac.jp

Abstract. Previously, we proposed an Auto-Regressive Hidden Markov Model (AR-HMM) and an accompanying parameter estimation method. An AR-HMM was obtained by combining an AR process with an HMM introduced as a non-stationary excitation model. We demonstrated that the AR-HMM can accurately estimate the characteristics of both articulatory systems and excitation signals from high-pitched speech. As the parameter estimation method iteratively executes learning processes of HMM parameters, the proposed method was calculation-intensive. Here, we propose two novel kinds of auto-regressive, non-stationary excited signal parameter estimation methods to reduce the amount of calculation required.

INTRODUCTION

The linear prediction (LP) method is widely used for the analysis of speech signals [1, 2]. However, several problems remain to be resolved. For example, (1) local peaks of LP spectral estimates are strongly biased toward harmonics, especially for high-pitched speech [3], and (2) addition of white noise to the Auto-Regressive (AR) process markedly alters the spectral estimate [4]. These phenomena result in deterioration of the perceived quality of re-synthesized speech and can also cause speech recognition errors.

LP methods assume that the excitation signal conforms to an Identically Independent Distributed (IID) normal distribution. However, the actual excitation signal indicates non-stationary properties especially in the case of high fundamental frequency. As a result, local peaks in the LP spectral envelope estimated from high-pitched speech are strongly biased toward harmonics. To correct this, we proposed an Auto-Regressive Hidden Markov

Model (AR-HMM) and an accompanying parameter estimation method [5] in which the HMM was introduced as a non-stationary excitation model. We also demonstrated that the proposed method can accurately estimate the characteristics of both articulatory systems and excitation signals from high-pitched speech.

As learning processes of HMM parameters are executed iteratively, the parameter estimation method proposed in our previous study is calculation-intensive. Here, we propose two novel auto-regressive, non-stationary excited signal parameter estimation methods to reduce the amounts of calculation required.

AUTO-REGRESSIVE HIDDEN MARKOV MODEL

Previously, we proposed an AR-HMM that was obtained by combining an AR process with an HMM introduced as a non-stationary excitation model. Figure 1 shows examples of this AR-HMM. The output probability distribution of each node in the excitation HMM is assumed to be a single normal distribution. The nodes of the first AR-HMM are concatenated in a ring state, so the state transition occurs in order. Therefore, this type of AR-HMM can be used to represent periodically excited signals. An ergodic HMM, as shown in the lower part of Fig. 1, can be used to represent an aperiodically excited signal. The AR-HMM can represent various types of signals through appropriate design of the network topology. The number of nodes and the prediction order are determined according to the signal. Usually, an Akaike Information Criterion (AIC) is employed to determine the model [6].

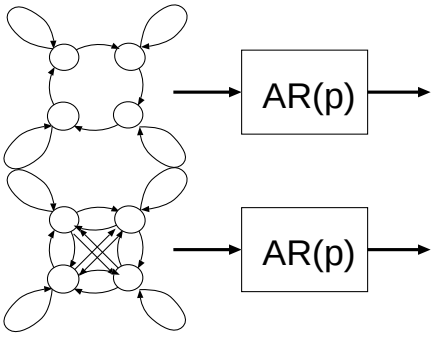

Figure 1: Examples of AR-HMMs.

ITERATIVE PARAMETER ESTIMATION METHOD USING VITERBI ALIGNMENT

The AR-HMM parameters are the AR coefficients and the parameters of the HMM. Previously, we presented an algorithm that iteratively estimates these

parameters from a signal $x(t), t = 0, \cdots, T-1$ [5]. In the following, P denotes the prediction order of the AR process. Let $\mathbf{a}^{(i)} = [a^{(i)}(1), \cdots, a^{(i)}(P)]^T$ represent the ith estimate of the AR coefficients. The ith estimate of the excitation signal $e^{(i)}(t), t = P, \cdots, T-1$ is given by:

$$\mathbf{e}_P^{(i)} = \mathbf{x}_P - \Omega \mathbf{a}^{(i)} \tag{1}$$

where

$$\mathbf{e}_P^{(i)} = [e^{(i)}(P)\ e^{(i)}(P+1)\ \cdots\ e^{(i)}(T-1)]^T \in R^{T-P},$$
$$\mathbf{x}_t = [x(t)\ x(t+1)\ \cdots\ x(t+T-P-1)]^T \in R^{T-P},$$
$$\Omega = [\mathbf{x}_{P-1}\ \mathbf{x}_{P-2}\ \cdots\ \mathbf{x}_0] \in R^{(T-P)\times P}$$

We allocate a unique number from $S = \{1, \cdots, N\}$ to each node of the excitation HMM to distinguish them from other nodes, where N is the number of nodes. Let $\mu_n^{(i)}, \sigma_n^{2\ (i)}, n \in S$ represent the ith estimates of the output distribution population parameters in each node. Given a state-transition sequence $s(t) \in S, t = P, \cdots, T-1$, the population parameters of an excitation signal at time t are given by $m^{(i)}(t) = \mu_{s(t)}^{(i)}, v^{(i)}(t) = \sigma_{s(t)}^{2\ (i)}$. Hence, the expectation vector of the excitation signal vector is represented by:

$$\mathbf{m}_P^{(i)} = [m^{(i)}(P)\ m^{(i)}(P+1)\ \cdots\ m^{(i)}(T-1)]^T \tag{2}$$

Based on the assumption that the samples of the excitation signal at different instants are mutually independent, the covariance matrix of the excitation signal vector is defined as a diagonal matrix given by:

$$\Sigma_P^{(i)} = \mathrm{diag}(v^{(i)}(P), v^{(i)}(P+1), \cdots, v^{(i)}(T-1)) \tag{3}$$

The algorithm for parameter estimation consists of the following processes.

1. The initial population parameters of the excitation signal are prepared as $\mathbf{m}_P^{(0)} = \mathbf{0}$, $\Sigma_P^{(0)} = \mathbf{I}$. Repeat the following processes from $i = 0$.

2. The AR coefficients $\mathbf{a}^{(i+1)}$ and the excitation signal $\mathbf{e}_P^{(i+1)}$ are estimated by maximizing the likelihood given by $L(\mathbf{e}_P^{(i+1)}; \mathbf{m}_P^{(i)}, \Sigma_P^{(i)})$.

3. The population parameters $\mathbf{m}_p^{(i+1)}$, $\Sigma_p^{(i+1)}$ of the excitation signal vector are estimated by maximizing the likelihood given by $L(\mathbf{e}_P^{(i+1)}; \mathbf{m}_P^{(i+1)}, \Sigma_p^{(i+1)})$.

4. If the likelihood has converged, the algorithm stops. Otherwise, repeat the above processes for $i \leftarrow i+1$ from step 2.

By repeating the above processes, the likelihood increases almost monotonically in practical situations and converges to the optimum or to a local optimum value.

The details of each step are as follows. In step 2, the AR coefficient vector can be obtained by solving

$$\frac{\partial}{\partial \mathbf{a}} \log L(\mathbf{x}_P - \Omega\mathbf{a}; \mathbf{m}_P^{(i)}, \Sigma_P^{(i)})\Big|_{\mathbf{a}=\mathbf{a}^{(i+1)}} = \mathbf{0}.$$

The solution is given by:

$$\mathbf{a}^{(i+1)} = [\Omega^T(\Sigma_P^{(i)})^{-1}\Omega]^{-1}\Omega^T(\Sigma_P^{(i)})^{-1}(\mathbf{x}_P - \mathbf{m}_P^{(i)}). \tag{4}$$

The excitation signal vector $\mathbf{e}_P^{(i+1)}$ is derived from (1).

In step 3, the population parameters of the excitation signal vector are estimated according to the following processes.

3.1 The Baum-Welch algorithm [7] estimates the population parameters $\mu_m^{(i+1)}, \sigma_m^{2\,(i+1)}, m \in S$ of each output distribution using $\mathbf{e}_P^{(i+1)}$.

3.2 The Viterbi algorithm [8] estimates a state transition sequence $s(t), t = P, P+1, \cdots, T-1$.

3.3 The expectation vector $\mathbf{m}_P^{(i+1)}$ and the diagonal covariance matrix $\Sigma_P^{(i+1)}$ of the excitation signal vector are estimated using (2) and (3).

PARAMETER ESTIMATION METHOD BASED ON THE EXPECTATION-MAXIMIZATION ALGORITHM

The method described in the previous section estimates AR coefficients and HMM parameters separately and iteratively; HMM parameter estimation requires the largest amount of calculation in the algorithm. When the method is applied to real-time processing, it is necessary to reduce the number of HMM parameter estimation iterations. To do this, we adopt an expectation-maximization (EM) algorithm. That is, using an EM algorithm, the AR coefficient estimation process can be embedded into the HMM parameter estimation process, and all the parameters can be estimated during a single HMM parameter estimation.

In the following, we describe the EM-based parameter estimation method for generalized situations where several signals are available for learning and the output distribution of each node is a Gaussian mixture. Let $x_n(t), n = 1, \ldots, N, t = 0, \ldots, T_n - 1$ and $a(k), k = 1, \ldots, P$ represent observed signals and AR coefficients, respectively. The excitation signal $e_n(t)$ emitted from the HMM is given by:

$$e_n(t) = x_n(t) + \sum_{k=1}^{P} a(k) x_n(t-k). \tag{5}$$

The output probability distribution $o_s(e)$ of the node s is a Gaussian mixture given by:

$$o_s(e) = \sum_{m=1}^{M} \lambda_{s,m} N(e; \mu_{s,m}, \sigma_{s,m}^2), \quad \sum_{m=1}^{M} \lambda_{s,m} = 1. \tag{6}$$

If we let π_s and q_{s_1,s_2} represent initial probabilities and transition probabilities, respectively, AR-HMM parameter θ is given by $\theta = \{a, \pi, q, \lambda, \mu, \sigma^2\}$. When observed signals $x_n(t)$, state-transition sequence $s_n(t)$ and Gaussian-distribution sequence $m_n(t)$ are given, the likelihood $L(\theta)$ of parameter θ is given by:

$$L(\theta \mid \mathbf{x}_n, \mathbf{s}_n, \mathbf{m}_n) = \pi_{s_n(0)} \prod_{t=0}^{T_n-1} q_{s_n(t),s_n(t+1)} \lambda_{s_n(t),m_n(t)} \times N(e_n(t); \mu_{s_n(t),m_n(t)}, \sigma^2_{s_n(t),m_n(t)}). \tag{7}$$

In the above equation, because the variables $s_n(t)$ and $m_n(t)$ are actually unobservable, it is necessary to estimate parameter θ from incomplete data $x_n(t)$. This can be achieved using an EM algorithm, which consists of two steps as follows.

- **E-Step**

  We represent the probability of the unobservable variables $\mathbf{s}_n, \mathbf{m}_n$ using the currently estimated parameter θ, then evaluate the expectation $Q(\hat{\theta} \mid \theta)$ of logarithmic likelihood $\log(L(\hat{\theta}))$ with respect to the unobservable variables, where $\hat{\theta}$ represents the updated parameter. This expectation is given according to the following equation:

$$Q(\hat{\theta} \mid \theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{T_n-1} \sum_{i \in S} \sum_{j \in S} \sum_{m=1}^{M} \gamma_n(t, i, j, m) \log(L(\hat{\theta})) \tag{8}$$

  where the parameter $\gamma_n(t, i, j, m)$ represents the probability that the mth Gaussian of the ith node emitted the excitation signal $e_n(t)$ and a transition from the ith node to the jth node occurred. This parameter is evaluated by:

$$\gamma_n(t, i, j, m) = \frac{\alpha(i, t-1) q_{i,j} \lambda_{i,m} N(e_n(t); \mu_{i,m}, \sigma^2_{i,m}) \beta(j, t)}{L(\theta \mid \mathbf{x}_n)} \tag{9}$$

  where α and β are obtained by applying a Forward-Backward algorithm to the excitation signal.

- **M-Step**

  In the M-step, the expectation $Q(\hat{\theta} \mid \theta)$ is maximized with respect to $\hat{\theta}$. The updated AR coefficients $\hat{a}(k)$ are given as a solution of $\partial Q / \partial \hat{a}(k) = 0, k = 1, \ldots, P$. The solution is represented by the following equation:

$$\begin{aligned} c_{k,l} &= \sum_{n=1}^{N} \sum_{t=0}^{T_n-1} \sum_{i \in S} \sum_{j \in S} \sum_{m=1}^{M} \gamma_n(t, i, j, m) \\ &\qquad \times x_n(t-k) x_n(t-l) / \sigma^2_{i,m} \\ d_k &= \sum_{n=1}^{N} \sum_{t=0}^{T_n-1} \sum_{i \in S} \sum_{j \in S} \sum_{m=1}^{M} \gamma_n(t, i, j, m) \end{aligned} \tag{10}$$

$$\times\{x_n(t)-\mu_{i,m}\}x_n(t-l)/\sigma_{i,m}^2$$

$$[\hat{a}(1),\cdots,\hat{a}(P)]^T=-\mathbf{C}^{-1}\mathbf{d}$$

where $\mathbf{C}=(c_{k,l})_{k,l=1,\cdots,P}$ and $\mathbf{d}=[d_1,\cdots,d_P]^T$.

The updated HMM parameters are given according to the following equations:

$$\hat{\pi}_s=\frac{\sum_{n=1}^{N}\sum_{j\in S}\sum_{m=1}^{M}\gamma_n(0,s,j,m)}{\sum_{n=1}^{N}\sum_{i\in S}\sum_{j\in S}\sum_{m=1}^{M}\gamma_n(0,i,j,m)} \tag{11}$$

$$\hat{q}_{s_1,s_2}=\frac{\sum_{n=1}^{N}\sum_{t=0}^{T_n-1}\sum_{m=1}^{M}\gamma_n(t,s_1,s_2,m)}{\sum_{n=1}^{N}\sum_{t=0}^{T_n-1}\sum_{j\in S}\sum_{m=1}^{M}\gamma_n(t,s_1,j,m)} \tag{12}$$

$$\hat{\lambda}_{s,m}=\frac{\sum_{n=1}^{N}\sum_{t=0}^{T_n-1}\sum_{j\in S}\gamma_n(t,s,j,m)}{\sum_{n=1}^{N}\sum_{t=0}^{T_n-1}\sum_{j\in S}\sum_{m'=1}^{M}\gamma_n(t,s,j,m')} \tag{13}$$

$$\hat{\mu}_{s,m}=\frac{\sum_{n=1}^{N}\sum_{t=0}^{T_n-1}\sum_{j\in S}\gamma_n(t,s,j,m)e_n(t)}{\sum_{n=1}^{N}\sum_{t=0}^{T_n-1}\sum_{j\in S}\gamma_n(t,s,j,m)} \tag{14}$$

$$\hat{\sigma}_{s,m}^2=\frac{\sum_{n=1}^{N}\sum_{t=0}^{T_n-1}\sum_{j\in S}\gamma_n(t,s,j,m)\left(e_n(t)-\mu_{s,m}\right)^2}{\sum_{n=1}^{N}\sum_{t=0}^{T_n-1}\sum_{j\in S}\gamma_n(t,s,j,m)} \tag{15}$$

The above two steps are iterated while replacing the current parameter θ with the updated parameter $\hat{\theta}$ until the likelihood converges.

POPULATION PARAMETER ESTIMATION BASED ON SAMPLE AVERAGE AND SAMPLE VARIANCE

The method described above can reduce the HMM parameter estimation iterations by adopting the EM algorithm. However, even if the HMM parameter estimation is executed only once, the estimation process still requires a large number of calculations. In this section, we propose a simplified method that does not use the HMM to estimate the population parameter of the excitation signal. In this method, the AR coefficients and the population parameters of the excitation signal are estimated separately and iteratively as in the method described in section 3, but the population parameters are evaluated with only the sample average and sample variance of the excitation signal.

The processes of this method are the same as processes 1 to 4 described in section 3. The difference is that the excitation signal vector population parameters $\mathbf{m}_p^{(i+1)}$, $\Sigma_p^{(i+1)}$ are estimated as follows. First, the sample average $m^{(i+1)}(t)$ and sample variance $v^{(i+1)}(t)$ are estimated with the following equations:

$$m^{(i+1)}(t)=\frac{1}{2T_s+1}\sum_{k=-T_s}^{T_s}e^{(i+1)}(t+k) \tag{16}$$

$$v^{(i+1)}(t) = \frac{1}{2T_s} \sum_{k=-T_s}^{T_s} \left(e^{(i+1)}(t+k) - m^{(i+1)}(t) \right)^2 \tag{17}$$

where $e^{(i+1)}(t)$ is the $(i+1)$th estimate of the excitation signal obtained from (1) and T_s is a parameter that determines the number of samples used for sample average and variance evaluations. The population parameters $\mathbf{m}_p^{(i+1)}$, $\Sigma_p^{(i+1)}$ are then obtained by substituting the evaluated sample average and variance into (2) and (3).

EXPERIMENTS

We conducted an experiment with synthetic speech to compare the estimation accuracy of the three proposed methods with the conventional autocorrelation linear prediction (ALP) method. In the following, we refer to the proposed methods described in Sections 3 to 5 as Method-1, Method-2, and Method-3, respectively. Speech was synthesized using the excitation signals of an impulse train and AR coefficients of order 16 extracted from a male's vowel /a/. The excitation signals were generated by adding white Gaussian noise $N(0, 0.1)$ to impulse trains of amplitude 50 in fundamental frequencies ranging from 100 Hz to 900 Hz. The sampling frequency was set to 16 kHz, the analysis frame length was set to 30 ms, and the prediction order was set to 16. The excitation HMM used for Method-1 and Method-2 consisted of two nodes: one representing an impulse, and another representing a noise segment. Both methods iterate the respective algorithm until the following condition is satisfied during 5 successive iterations: $|\frac{\log(L(\hat{\theta}))-\log(L(\theta))}{\log(L(\theta))}| < 1.0e-7$. The parameter T_s used for Method-3 was set to 5. The estimated AR coefficients were transformed to LPC Mel-Cepstrum coefficients, which play a very important role in applications such as speech recognition. The estimation accuracy of each method was then evaluated by the Euclidean distance between the original LPC Mel-Cepstrum coefficients and those obtained by this method.

Figure 2 shows the vocal tract spectra estimated by Method-1, Method-2, Method-3, and the ALP method. The original vocal tract spectrum that was used for synthesizing speech signals is shown at the top of each figure. Figure 3 shows the evaluated Mel-Cepstrum distances of all of the methods. As shown in these figures, the estimation accuracy of the ALP tended to degrade as the fundamental frequency increased. In contrast, Method-1 can extract speech features in the widest range of fundamental frequency; its estimation accuracy was also the highest of all of the methods examined. The estimation accuracy of Method-2 up to a fundamental frequency of 550 Hz was almost equivalent to that of Method-1. In the fundamental frequency range above 600 Hz, however, the estimation accuracy of Method-2 deteriorated markedly as compared to the other methods. In Method-3, the estimation accuracy tended to asymptotically approach that of the ALP when the fundamental frequency exceeded 600 Hz. However, the estimation accuracy was very close

to those of Method-1 and Method-2 up to a fundamental frequency of 550 Hz. Figure 4 shows the processing times required by each method given as values relative to the processing time of Method-1.

These results indicated that all of the proposed methods are capable of extracting features when analyzing speech signals with fundamental frequencies below 550 Hz. In addition, Method-3 can be applied to real-time processing because it does not use the HMM to estimate the population parameter of the excitation signal. However, Method-1 is still the best method if maximum estimation accuracy is necessary and there is sufficient time.

CONCLUSIONS

Here, we proposed two novel auto-regressive, non-stationary excited signal parameter estimation methods, in addition to the method proposed previously. We are currently planning to construct a singing voice recognition system and apply the proposed method to feature extraction at the front-end of the system.

REFERENCES

[1] F.Itakura and S.Saito, "A statistical method for estimation of speech spectral density and formant frequencies," Electronics and Communications in Japan, Vol.53-A, No.1, pp.36-43, January 1970.

[2] B.S.Atal and S.L.Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., Vol.50, pp.637-644, 1971.

[3] J.Makhoul, "Linear Prediction: A Tutorial Review," in Proc.of IEEE, Vol.63, No.4, pp.561-580, April 1975.

[4] S.M.Kay, "The Effects of Noise on the Autoregressive Spectral Estimator," IEEE ASSP-27, No.5, pp.478-485, Oct. 1979.

[5] A.Sasou, K.Tanaka, "Glottal Source Modeling Using HMM and Robust Analysis of High Fundamental Frequency Speech," in Trans. IEICE, Vol. J84-D-II, No.9, pp.1960-1969, Sep. 2001(in Japanese).

[6] H.Akaike, "A new-look at the statistical model identification," IEEE Trans. Autom. Control, Vol.AC-19, No.6, pp.716-723, 1974.

[7] L.E.Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic function of a Markov process," Inequalities, Vol.3, pp.1-8, 1972.

[8] A.J.Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Trans. on Information Theory, Vol.IT-13, No.2, pp.260-269, April 1967.

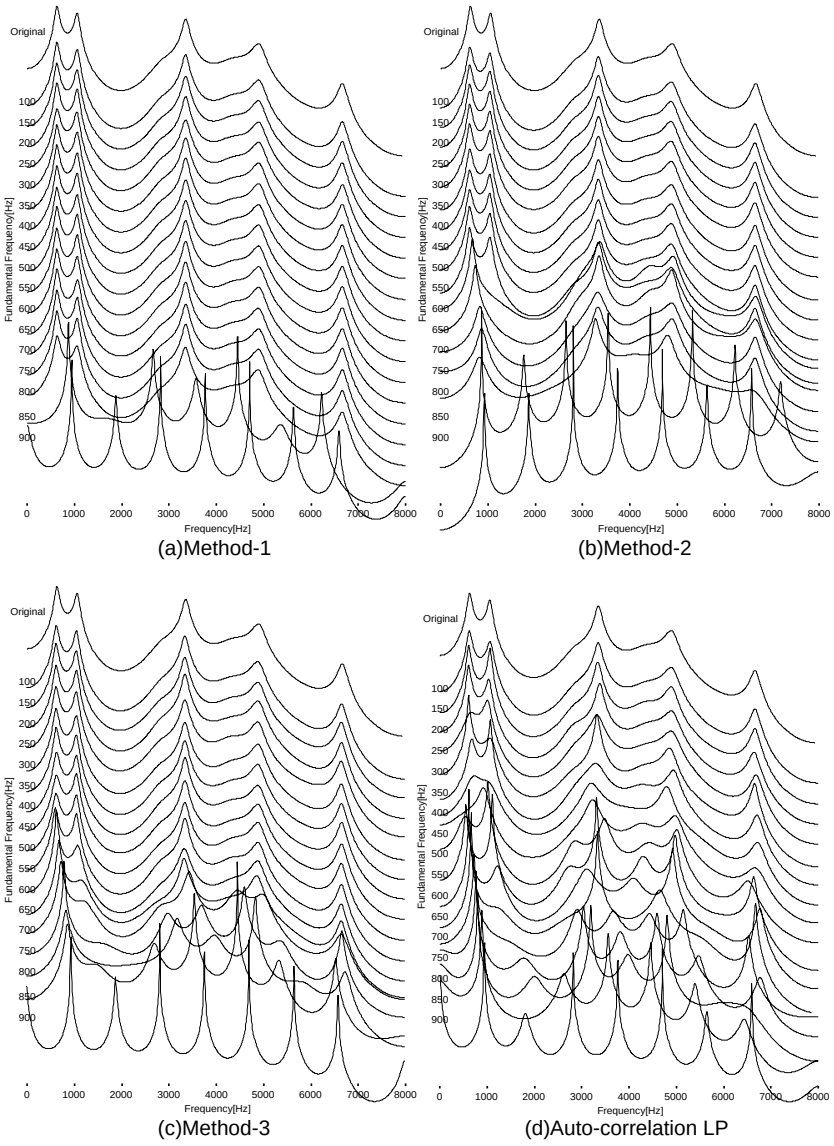

Figure 2: Vocal tract spectra estimated by (a)Method-1, (b)Method-2, (c)Method-3 and (d)Auto-Correlation LP

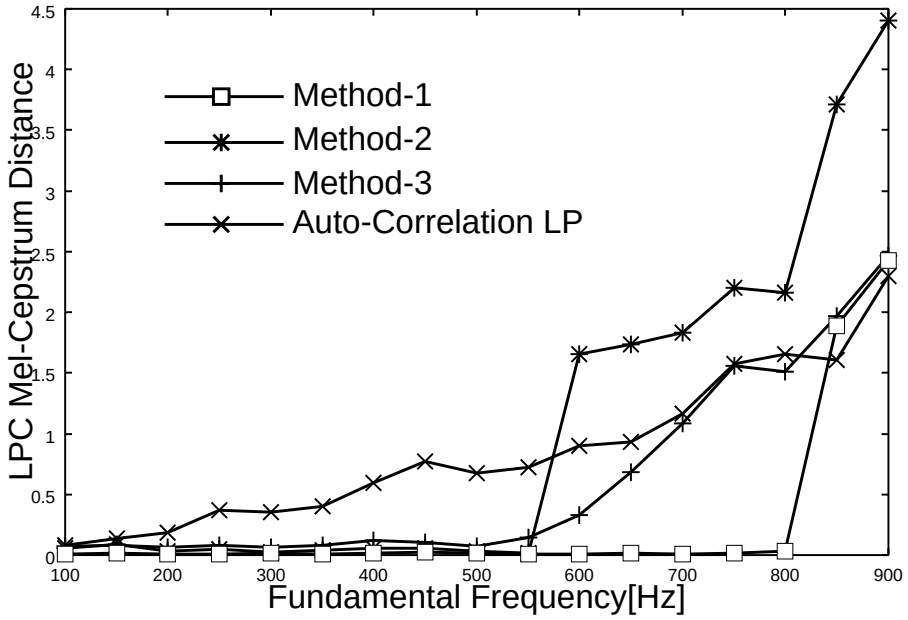

Figure 3: Evaluated Mel-Cepstrum Distances

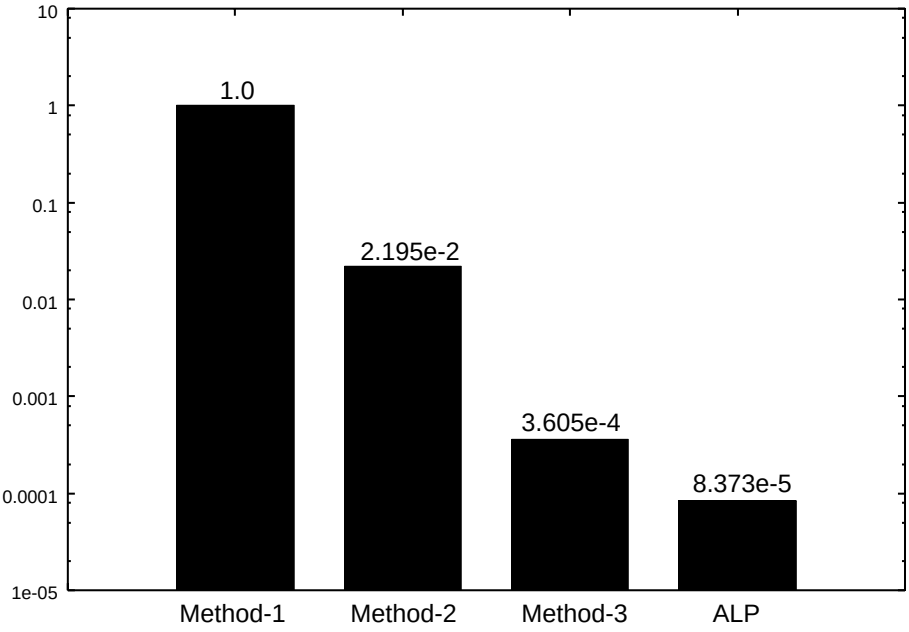

Figure 4: Relative processing times