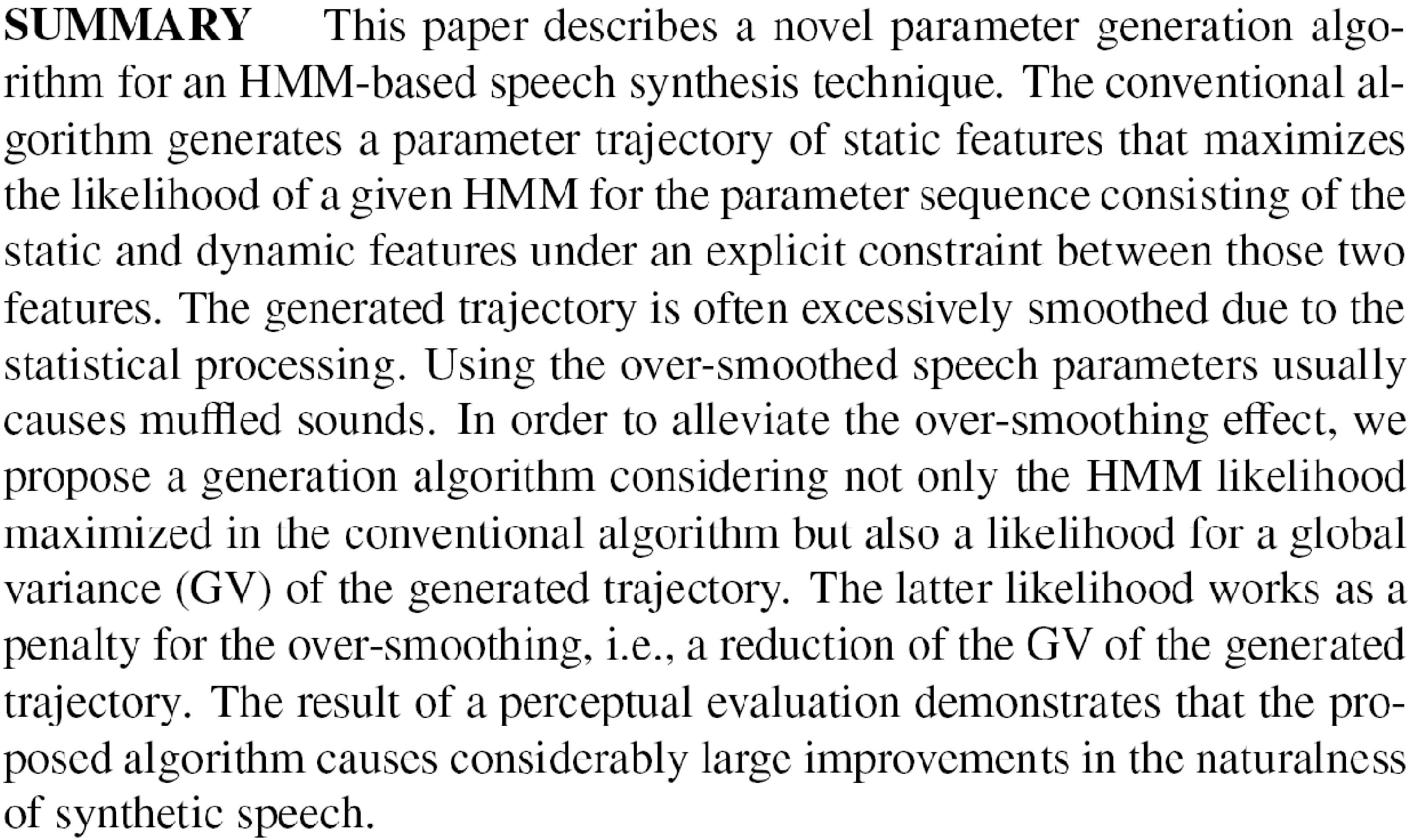
A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis

Source: IEICE TRANS. INF. & SYST., VOL.E90–D, NO.5 MAY 2007

Author : Tomoki TODA *and* Keiichi TOKUDA

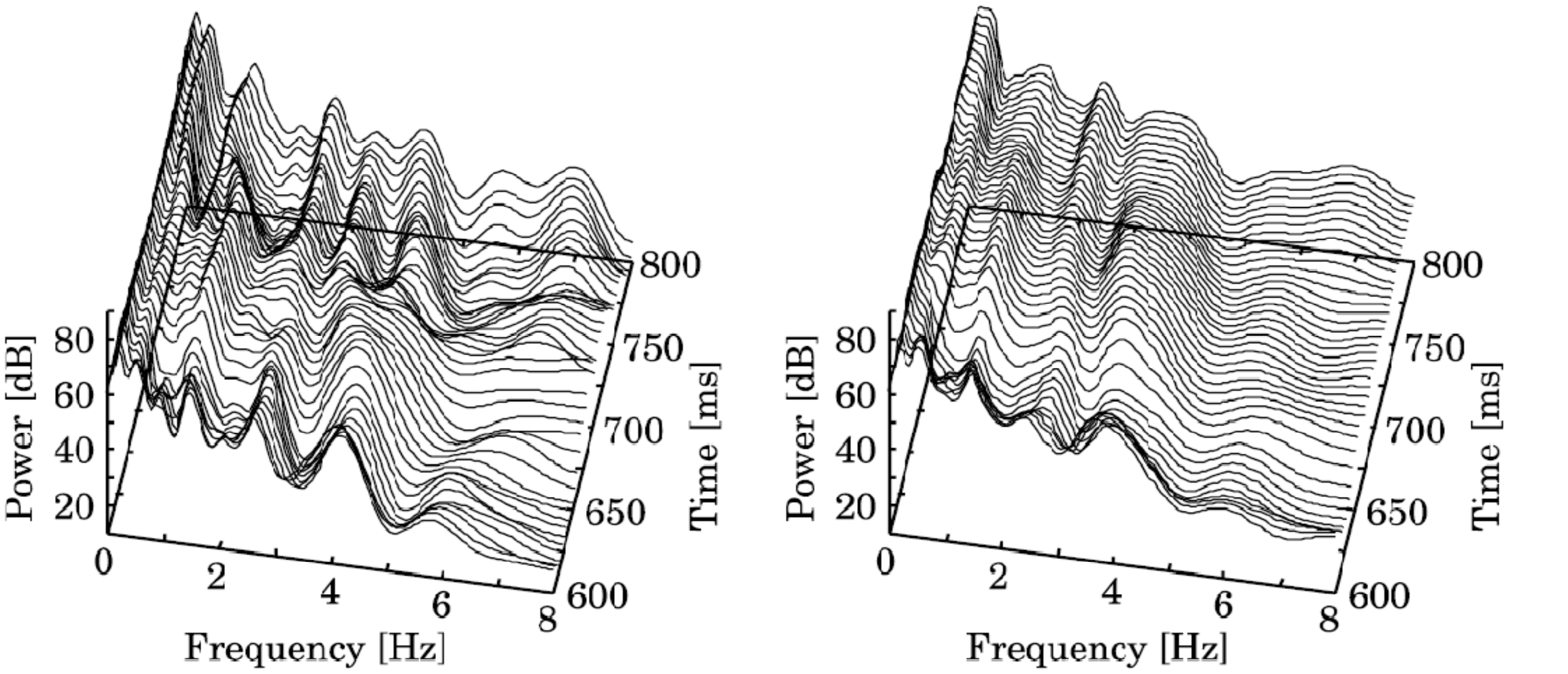
Professor : 陳嘉平

Reporter : 楊治鏞



SUMMARY This paper describes a novel parameter generation algorithm for an HMM-based speech synthesis technique. The conventional algorithm generates a parameter trajectory of static features that maximizes the likelihood of a given HMM for the parameter sequence consisting of the static and dynamic features under an explicit constraint between those two features. The generated trajectory is often excessively smoothed due to the statistical processing. Using the over-smoothed speech parameters usually causes muffled sounds. In order to alleviate the over-smoothing effect, we propose a generation algorithm considering not only the HMM likelihood maximized in the conventional algorithm but also a likelihood for a global variance (GV) of the generated trajectory. The latter likelihood works as a penalty for the over-smoothing, i.e., a reduction of the GV of the generated trajectory. The result of a perceptual evaluation demonstrates that the proposed algorithm causes considerably large improvements in the naturalness of synthetic speech.

Over-Smoothing of Generated Parameters



Natural

Generated

Fig. 1 An example of natural and generated spectral segments for a phoneme sequence "a-/n a u/+n."

Conventional Parameter Generation Algorithm

- D-dimensional static feature vector at frame t.

$$\boldsymbol{c}_t = [c_t(1), c_t(2), \cdots, c_t(d), \cdots, c_t(D)]^\top$$

- Speech parameter vector

$$\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, \Delta^{(2)}\boldsymbol{c}_t^\top]^\top$$

- Dynamic feature vectors $\Delta^{(1)}c_t$, $\Delta^{(2)}c_t$, which are calculated by

$$\Delta^{(n)}c_t = \sum_{\tau=-L_-^{(n)}}^{L_+^{(n)}} w^{(n)}(\tau)\, c_{t+\tau}, \quad n = 1, 2$$

Conventional Parameter Generation Algorithm

- Parameter vectors at all frames over an utterance is regarded as a time sequence vector.

$$\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \cdots, \boldsymbol{o}_t^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top,$$

$$\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top,$$

- The relationship between those two vectors is represented as

$$o = Wc$$

where W is the 3*DT*-by-*DT* matrix

$$W = \left[W_1, W_2, \cdots, W_t, \cdots, W_T\right]^\top$$

$$W_t = \left[w_t^{(0)}, w_t^{(1)}, w_t^{(2)}\right],$$

$$w_t^{(n)} = \Bigg[\overset{1st}{0}, \cdots, 0, \overset{(t-L_-^{(n)})\text{-th}}{w^{(n)}(-L_-^{(n)})}, \cdots, \overset{(t)\text{-th}}{w^{(n)}(0)}, \cdots,$$

$$\overset{(t-L_+^{(n)})\text{-th}}{w^{(n)}(-L_+^{(n)})}, 0, \cdots, \overset{T\text{-th}}{0}\Bigg]^\top, \qquad n = 0, 1, 2$$

$$L_-^{(0)} = L_+^{(0)} = 0, \text{ and } w^{(0)}(0) = 1.$$

HMM Likelihood

- A likelihood of a given continuous mixture HMM λ for the parameter sequence vector o is written as

$$P(o \mid \lambda) = \sum_{\text{all } Q} P(o, Q \mid \lambda)$$

where

$$Q = \{(q_1, i_1), (q_2, i_2), \cdots (q_T, i_T)\}$$

is the state and mixture sequence, i.e., (q, i) indicates the *i*-th mixture of state *q*.

Parameter Sequence Generation Based on Maximum Likelihood Criterion

- We determine the parameter sequence of static features c that maximizes the HMM likelihood.
- In order to reduce computation cost, the current HMM-based speech synthesis system determines the sub-optimum state sequence $\hat{q} = \{\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_t, \cdots, \hat{q}_T\}$ independently of o as follows

$$\hat{q} = \arg\max P(q \mid \lambda)$$

where $P(q \mid \lambda)$ is a likelihood of the state duration model.

Parameter Sequence Generation Based on Maximum Likelihood Criterion

- We maximize the following log-scaled likelihood with respect to c ,

$$\log P(\boldsymbol{o}|\hat{\boldsymbol{Q}}, \lambda) = -\frac{1}{2}\boldsymbol{o}^\top \hat{\boldsymbol{U}}^{-1}\boldsymbol{o} + \boldsymbol{o}^\top \hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}} + K,$$

$$\hat{\boldsymbol{\mu}} = \left[\boldsymbol{\mu}_{\hat{q}_1,\hat{i}_1}^\top, \boldsymbol{\mu}_{\hat{q}_2,\hat{i}_2}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_t,\hat{i}_t}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_T,\hat{i}_T}^\top\right]^\top,$$

$$\hat{\boldsymbol{U}}^{-1} = \mathrm{diag}\left[\boldsymbol{U}_{\hat{q}_1,\hat{i}_1}^{-1}, \boldsymbol{U}_{\hat{q}_2,\hat{i}_2}^{-1}, \cdots, \boldsymbol{U}_{\hat{q}_t,\hat{i}_t}^{-1}, \cdots, \boldsymbol{U}_{\hat{q}_T,\hat{i}_T}^{-1}\right],$$

Parameter Sequence Generation Based on Maximum Likelihood Criterion

$$\frac{\partial \log P(Wc \mid \hat{Q}, \lambda)}{\partial c} = 0$$

$$c = \left(W^{\top} \hat{U}^{-1} W \right)^{-1} W^{\top} \hat{U}^{-1} \hat{\mu}.$$

- As shown in the above equation, this algorithm is not a frame-based process but a trajectory-based one, i.e., simultaneously generating static vectors at all frames.

Global Variance

- A GV of the static features over a time sequence is calculated by

$$v(c) = [v(1), v(2), \cdots, v(d), \cdots, v(D)]^{\top}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(c_t(d) - \overline{c}(d)\right)^2,$$

$$\overline{c}(d) = \frac{1}{T}\sum_{\tau=1}^{T} c_{\tau}(d).$$

- This paper calculates the GV utterance by utterance.

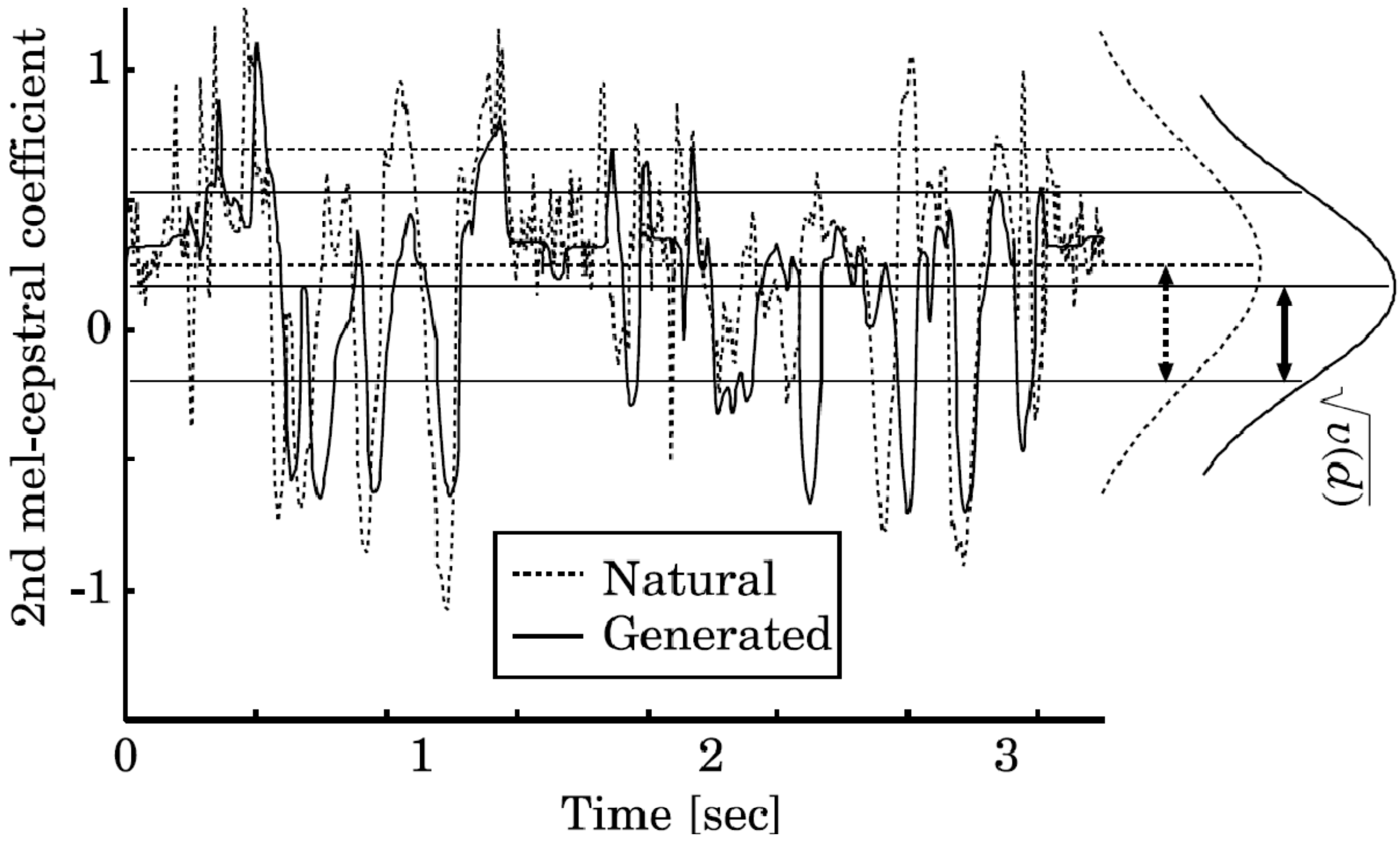


Fig. 2 Natural and generated mel-cepstrum sequences. A square root of the GV of each sequence is shown as a bidirectional arrow.

Proposed Likelihood

- The proposed method considers not only the HMM likelihood for the static and dynamic feature vectors but also the likelihood on the GV.

$$P(o \mid \lambda, \lambda_v) = \sum_{\text{all } Q} P(o, Q \mid \lambda)^{\omega} \, P(v(c) \mid \lambda_v),$$

where $P(v(c) \mid \lambda_v)$ is modeled by a single Gaussian distribution. The constant ω denotes the weight for controlling a balance between the two likelihoods.

Parameter Sequence Generation Considering GV Based on Maximum Likelihood Criterion

- Note that the proposed likelihood is a function of c .
- The following log-scaled likelihood is maximized with respect to c under the condition of determined $\hat{Q}$.

$$
\begin{aligned}
L &= \log\left[P(\boldsymbol{o}|\hat{\boldsymbol{Q}}, \lambda)^{\omega} P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{\lambda}_v)\right] \\
&= \omega\left(-\frac{1}{2}\boldsymbol{c}^{\top}\boldsymbol{W}^{\top}\hat{\boldsymbol{U}}^{-1}\boldsymbol{W}\boldsymbol{c} + \boldsymbol{c}^{\top}\boldsymbol{W}^{\top}\hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}}\right) \\
&\quad -\frac{1}{2}\boldsymbol{v}(\boldsymbol{c})^{\top}\boldsymbol{U}_v^{-1}\boldsymbol{v}(\boldsymbol{c}) + \boldsymbol{v}(\boldsymbol{c})\boldsymbol{U}_v^{-1}\boldsymbol{\mu}_v + \overline{K},
\end{aligned}
$$

Parameter Sequence Generation Considering GV Based on Maximum Likelihood Criterion

- To determine c, we iteratively update c with the gradient method,

$$c^{(i+1)\text{-th}} = c^{(i)\text{-th}} + \alpha \cdot \delta c^{(i)\text{-th}}$$

- Steepest decent algorithm:

$$\delta \boldsymbol{c}^{(i)\text{-th}} = \left.\frac{\partial L}{\partial \boldsymbol{c}}\right|_{\boldsymbol{c}=\boldsymbol{c}^{(i)\text{-th}}}.$$

- Newton-Raphson method:

$$\delta \boldsymbol{c}^{(i)\text{-th}} = -\left(\frac{\partial^2 L}{\partial \boldsymbol{c} \partial \boldsymbol{c}^\top}\right)^{-1} \left.\frac{\partial L}{\partial \boldsymbol{c}}\right|_{\boldsymbol{c}=\boldsymbol{c}^{(i)\text{-th}}}.$$

Steepest Decent Algorithm

- The first derivative is calculated by

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{c}} &= \omega\left(-\boldsymbol{W}^\top \hat{\boldsymbol{U}}^{-1}\boldsymbol{W}\boldsymbol{c} + \boldsymbol{W}^\top \hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}}\right) \\
&\quad + \left[\boldsymbol{v}_1'^\top, \boldsymbol{v}_2'^\top, \cdots, \boldsymbol{v}_t'^\top, \cdots, \boldsymbol{v}_T'^\top\right]^\top, \\
\boldsymbol{v}_t' &= \left[v_t'(1), v_t'(2), \cdots, v_t'(d), \cdots, v_t'(D)\right]^\top, \\
v_t'(d) &= -\frac{2}{T}\left(c_t(d) - \overline{c}(d)\right)\boldsymbol{p}_v^{(d)\top}\left(\boldsymbol{v}(\boldsymbol{c}) - \boldsymbol{\mu}_v\right),
\end{aligned}
$$

- where $\boldsymbol{p}_v^{(d)}$ is the *d*-th column vector of $\boldsymbol{P}_v = \boldsymbol{U}_v^{-1}$.

Newton-Raphson method

$$
\begin{aligned}
\frac{\partial^2 L}{\partial \boldsymbol{c} \partial \boldsymbol{c}^\top} &\simeq -\omega \cdot \mathrm{diag}\left[\boldsymbol{r}^\top + \boldsymbol{v}''^\top\right], \\
\boldsymbol{r} &= \left[r_1, r_2, \cdots, r_t, \cdots, r_T\right]^\top, \\
\boldsymbol{v}'' &= \left[v''_1, v''_2, \cdots, v''_t, \cdots, v''_T\right]^\top, \\
\boldsymbol{r}_t &= \left[r_t(1), r_t(2), \cdots, r_t(d), \cdots, r_t(D)\right]^\top, \\
\boldsymbol{v}''_t &= \left[v''_t(1), v''_t(2), \cdots, v''_t(d), \cdots, v''_t(D)\right]^\top, \\
r_t(d) &= \boldsymbol{w}^{((t-1)D+d)^\top} \hat{\boldsymbol{U}}^{-1} \boldsymbol{w}^{((t-1)D+d)}, \\
v''_t(d) &= -\frac{2}{T^2}\left\{(T-1)\boldsymbol{p}_v^{(d)^\top}\left(\boldsymbol{v}(\boldsymbol{c}) - \boldsymbol{\mu}_v\right)\right. \\
&\qquad \left. + 2p_v^{(d)}(d)\left(c_t(d) - \overline{c}(d)\right)^2\right\},
\end{aligned}
$$

Gradient Method

- There are mainly two settings of the initial trajectory $c^{(0)\text{-th}}$.
  - □ One is to use the conventional trajectory.
  - □ The other is to use the trajectory c' linearly converted from the conventional one as follows

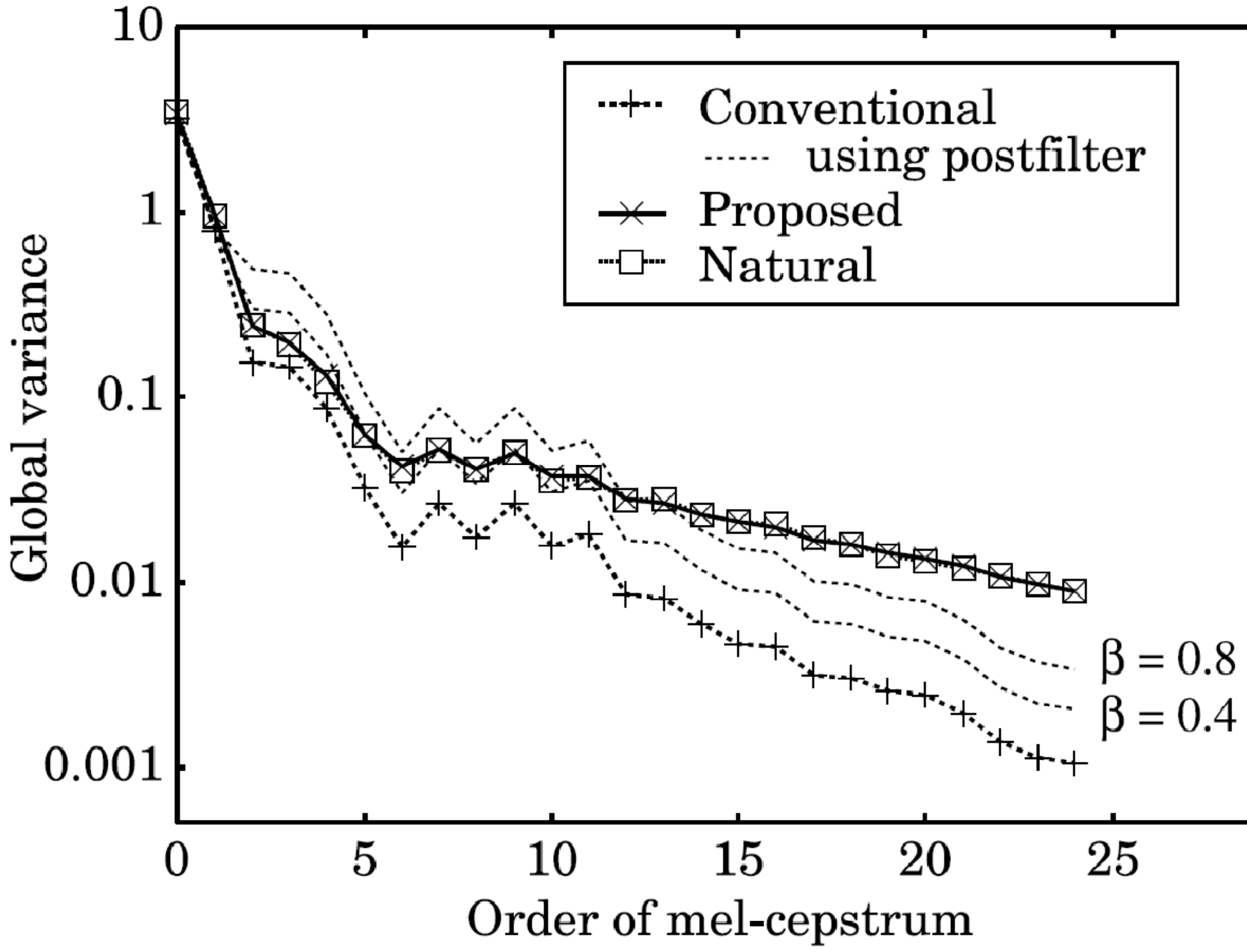
$$c_t'(d) = \sqrt{\frac{\mu_v(d)}{v(d)}}\left(c_t(d) - \overline{c}(d)\right) + \overline{c}(d)$$

Experimental Conditions

- ATR Japanese speech database
  - Two males, two female
  - 450 sentence

- As a spectral parameter, we used 0th through 24th mel-cepstral coefficients obtained from the smoothed spectrum analyzed by STRAIGHT.

Objective Evaluations



Perceptual Evaluation

- We conducted an opinion test on the naturalness of synthetic speech to demonstrate the effectiveness of the proposed method.
- Seven Japanese listeners participated in the test.
- Each listener evaluated 25 samples consisting of five sentences for each speaker.

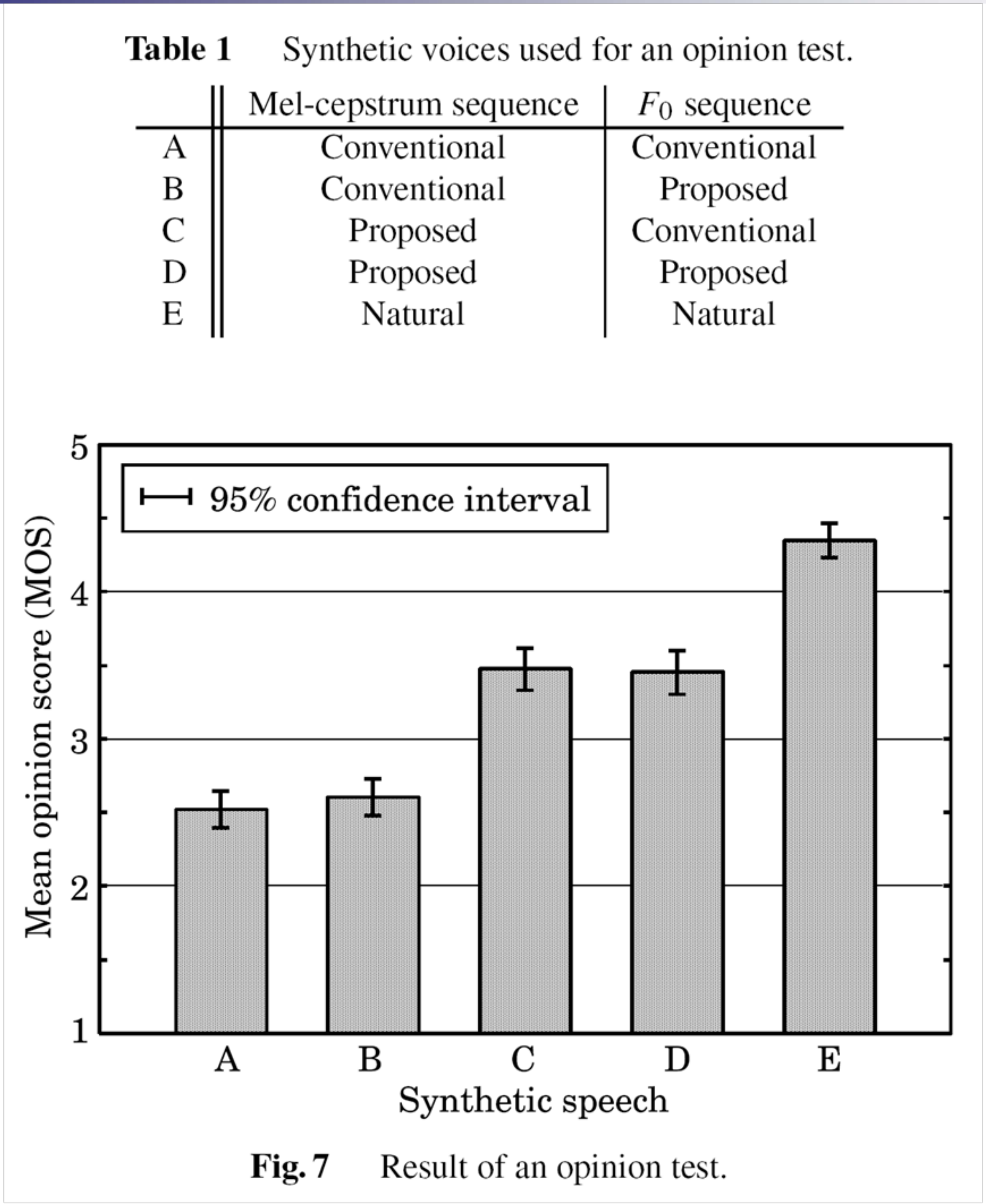


Table 1 Synthetic voices used for an opinion test.

	Mel-cepstrum sequence	F_0 sequence
A	Conventional	Conventional
B	Conventional	Proposed
C	Proposed	Conventional
D	Proposed	Proposed
E	Natural	Natural

Fig. 7 Result of an opinion test.

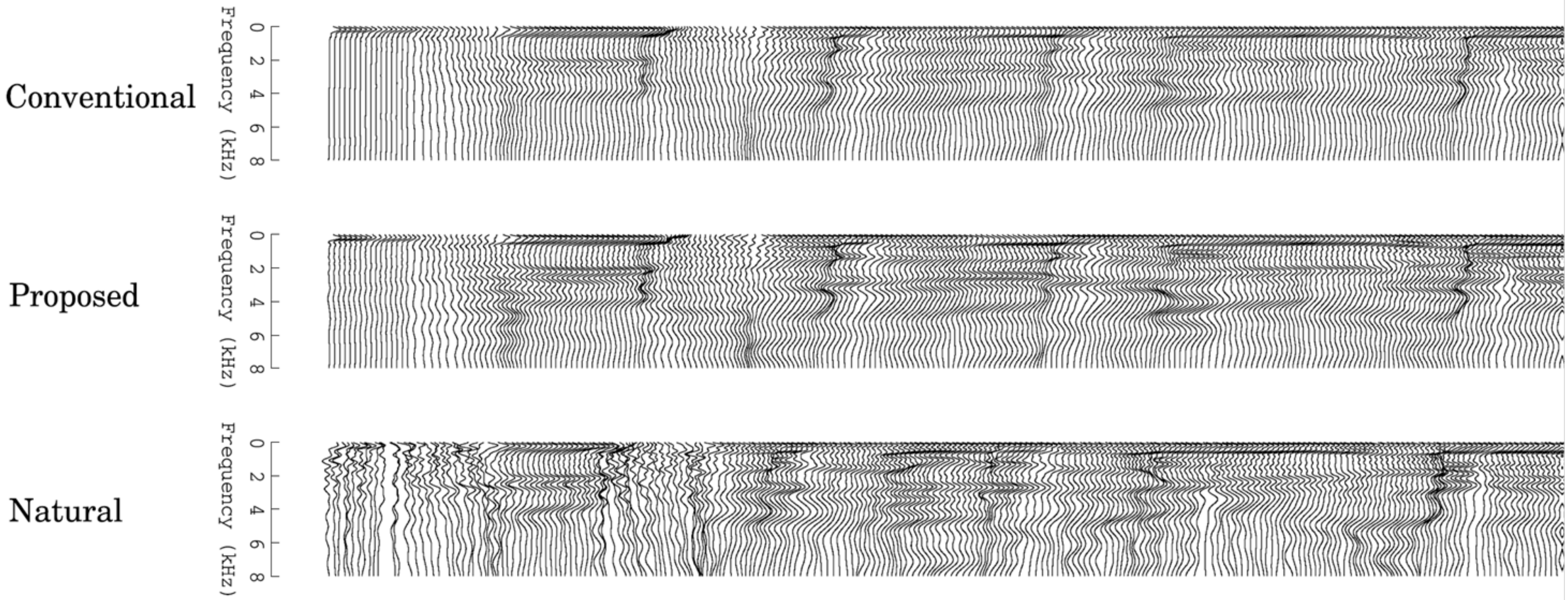


Fig. 8 An example of spectrum sequences of generated speech with conventional algorithm, generated speech with proposed algorithm, and natural speech. Note that phoneme duration of the natural sequence is different from those of the generated ones.