IEICE TRANS. INF. & SYST., VOL.E94–D, NO.10 OCTOBER 2011

PAPER

Probabilistic Concatenation Modeling for Corpus-Based Speech Synthesis

Shinsuke SAKAI[†a)], **Tatsuya KAWAHARA**[†], ***and*** **Hisashi KAWAI**[††], ***Members***

SUMMARY The measure of the goodness, or inversely the cost, of concatenating synthesis units plays an important role in concatenative speech synthesis. In this paper, we present a probabilistic approach to concatenation modeling in which the goodness of concatenation is measured by the conditional probability of observing the spectral shape of the current candidate unit given the previous unit and the current phonetic context. This conditional probability is modeled by a conditional Gaussian density whose mean vector has a form of linear transform of the past spectral shape. Decision tree-based parameter tying is performed to achieve robust training that balances between model complexity and the amount of training data available. The concatenation models are implemented for a corpus-based speech synthesizer, and the effectiveness of the proposed method was confirmed by an objective evaluation as well as a subjective listening test. We also demonstrate that the proposed method generalizes some popular conventional methods in that those methods can be derived as the special cases of the proposed method.

key words: *speech synthesis, unit selection, concatenation cost, join cost*

1. Introduction

The corpus-based unit concatenation approach to speech synthesis has been widely explored in the research community in recent years [1]–[6]. In this approach, an optimal sequence of subphone, phone, or non-uniform (e.g. [1], [2]) synthesis units are chosen from a large inventory of units to synthesize speech from the input text through the minimization of the overall cost. The overall cost is often modeled as the weighted sum of target costs and concatenation (or join) costs defined over various features of synthesis units such as spectral shape, intonation contour, and segmental duration. In order to achieve as smooth concatenation of successive units as possible, various approaches to concatenation costs have been explored based on distance or distortion of acoustic parameters such as F_0, power, and cepstrum [2], [3], often combined with empirical table-driven costs looked up by phonetic or prosodic features [4], [5], [7]. There is also an approach to use normalized cross-correlation at the waveform level to represent the concatenation cost [6]. Since establishing a good model of concatenation cost is one of the most important issues that influence the quality of concatenative speech synthesis, there has also been a number of research efforts focused on the issue of concatenation cost [8]–[12], in which various spectral feature parameters and distance measures are investigated and compared.

In our approach to concatenation modeling proposed in this paper, we depart from the traditional view of concatenation cost based on "distance" or "distortion" and take a probabilistic view of concatenation cost where concatenation modeling is done with a probabilistic model that captures how likely it is to observe the spectral shape of the current unit given the spectral shape of the previous unit within the current linguistic context. For the modeling of this conditional probability, we make use of *conditional Gaussian* models. The mean vector of a conditional Gaussian density has a form of linear transform of some other vector, which is useful for representing the correlation between two random variables. An example of the use of conditional Gaussian in speech processing is found in *autoregressive HMMs* [13], where the observation vector from a state is conditioned not only on the identity of the current state but also on the observation from the previous state.

In order to perform as fine modeling as possible with a limited amount of training data, the parameters in the conditional Gaussian models for various different contexts are tied using tree-based context clustering. We show that this clustering can be done efficiently using the sufficient statistics appearing in the maximum likelihood estimate that we derive from the definition of conditional Gaussian density.

The effectiveness of the proposed approach to concatenation modeling is demonstrated through comparative objective and subjective evaluation experiments using a unit selection-based concatenative speech synthesis system [14]. In this speech synthesis system, spectral target costs are given by multivariate Gaussian densities and F_0 target costs are computed using statistical additive F_0 models [15], [16]. The duration target costs are given by scalar Gaussian models identified by phonetic features and positional features related to syllable, word, and phrase units as well as features related to prosodic events such as lexical stress and pitch accent [14]. The proposed concatenation modeling method as well as a conventional method of concatenation cost are employed in this system and compared in the experiment.

The rest of the paper is organized as follows. Section 2 gives an overview of the probabilistic concatenation models with a mathematical definition and the derivation of maximum likelihood estimate from the model definition. A robust and efficient training method for the models based on phonetic decision tree-based context tying is described in Sect. 3. In Sect. 4, we discuss the mathematical relationships

Manuscript received April 27, 2011.

[†]The authors are with Academic Center for Computing and Media Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.

[††]The author is with Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Kyoto-fu, 619–0289 Japan.

a) E-mail: sakai@ar.media.kyoto-u.ac.jp

DOI: 10.1587/transinf.E94.D.2006

of the proposed method with other conventional approaches. Experimental results are presented in Sect. 5 where we examine how linear transforms for conditional means of the models are trained from the corpus. Objective and subjective evaluation results are also presented. The last section presents our conclusion.

2. Probabilistic Concatenation Models

We model the goodness of concatenation of the spectral characteristics of the adjacent synthesis units in terms of the conditional probability $P(\boldsymbol{h}(u_k)|\boldsymbol{t}(u_{k-1}), c_k)$ of observing the feature vector $\boldsymbol{h}(u_k)$ that represents the spectral characteristics of the initial part (or *head*) of the current unit u_k given the feature vector $\boldsymbol{t}(u_{k-1})$ that represents the spectral characteristics of the ending part (or *tail*) of the previous unit u_{k-1} and the context c_k for the current unit. These feature vectors $\boldsymbol{h}(u)$ for the head and $\boldsymbol{t}(u)$ for the tail of the unit u can be implemented as the average of spectral features over a few frames at the beginning and the end of the unit, respectively, for example. This concatenation probability is modeled by a conditional Gaussian density,

$$P(\boldsymbol{h}(u_k)|\boldsymbol{t}(u_{k-1}), c_k) = \mathcal{N}(\boldsymbol{h}(u_k)|\boldsymbol{B}_{c_k} \cdot \boldsymbol{t}(u_{k-1}) + \boldsymbol{b}_{c_k}, \boldsymbol{\Sigma}_{c_k}), \quad (1)$$

where $\boldsymbol{h}(u_k)$ and $\boldsymbol{t}(u_{k-1})$ are d-dimensional vectors, $\boldsymbol{B}_{c_k}$ is a $d \times d$ regression matrix with the j-th row representing regression coefficients for the j-th component of $\boldsymbol{h}(u_k)$, $\boldsymbol{b}_{c_k}$ is a d-dimensional vector of intercepts, and $\boldsymbol{\Sigma}_{c_k}$ is a $d \times d$ covariance matrix. The context c_k may consist of information such as the identities of the current and preceding phones and the position of the current unit in the current phone if the granularity of the unit is smaller than a phone. We drop the suffix c_k for simplicity of notation hereafter. A graphical model representation of this conditional Gaussian model is depicted in Fig. 1.

The use of conditional Gaussian is motivated by the observation that the spectral shapes, i.e. the energy distribution patterns in the frequency domain, across segment boundaries are not always similar between both sides of the boundary when we examine them after labeling the natural speech to segment it into small units such as phone, half phone, or HMM states. When we see a boundary between some two vowels, for example, we observe that the spectral shapes are almost same across the boundary, whereas an abrupt spectral change is observed at the boundary and the spectral shapes are totally different across the boundary between some obstruent consonant and a vowel. In some boundaries such as those between a voiced consonant and a vowel, we see not only an abrupt change but also a continuity or a some kind of dependency between both sides of the boundary. Figure 2 presents conceptual graphs that depict the relationships between the tail and the head of two consecutive units in two extreme cases, using i-th components of hypothetical spectral feature vectors. Figure 2 (a) corresponds to a case where the i-th feature vector components are very similar across the unit boundary, e.g. a case where a vowel is followed by the same vowel. In this kind of situation, the i-th row of the matrix $\boldsymbol{B}$ that contributes to the transform of the i-th component is considered to be dominated by the (i, i)-component and the i-th component of the constant vector $\boldsymbol{b}$ is close to zero. On the other hand, if there is a case like Fig. 2 (b), where $h_i(u_k)$, the i-th component of the head feature vector of the current unit, almost has the constant value, the i-th row of the regression matrix $\boldsymbol{B}$ is considered to be close to zero vector and the i-th element of the vector $\boldsymbol{b}$ will be the significant contributor to the i-th component of the conditional mean vector. In general, the i-th row of $\boldsymbol{B}$ should have some meaningful values in multiple columns, if $h_i(u_k)$ has dependencies to i-th and/or some other components of $\boldsymbol{t}(u_k)$. With the characteristics of the conditional Gaussian models described above, we can expect a properly high score if adjacent units have a typical continuity and/or discontinuity patterns in their tail and head feature vectors, irrespective of the similarity of these two vectors.

In general, the proposed method is considered to be applicable to the approaches with uniform sized units such as phone, half-phone, and HMM state, where all the competing hypotheses have the same number of concatenation points. In the current paper, we specifically demonstrate the effectiveness of the proposed method through the implementation

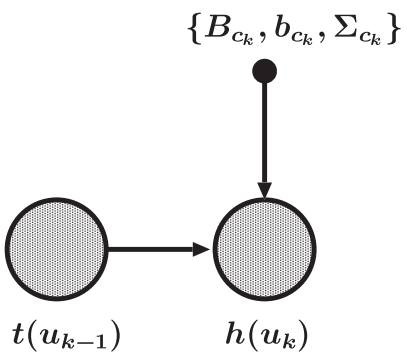


Fig. 1 A graphical model representation of the conditional Gaussian concatenation model.

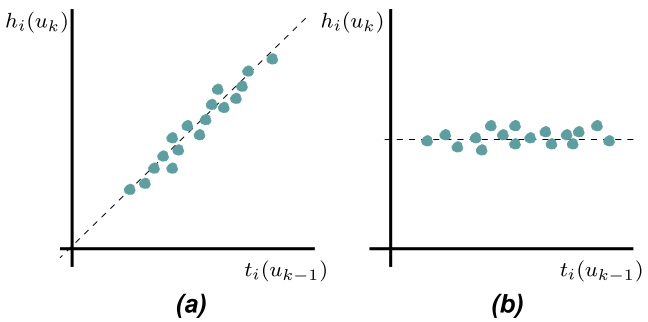


Fig. 2 Schematic diagram representing the relationship between $h_i(u_k)$ and $t_i(u_{k-1})$, which are the i-th components of the vectors $\boldsymbol{h}(u_k)$ and $\boldsymbol{t}(u_{k-1})$, in two extreme cases. (a) $h_i(u_k)$ is very similar to $t_i(u_{k-1})$. (b) $h_i(u_k)$ is almost independent of $t_i(u_{k-1})$.

of speech synthesis system with phone sized units.

2.1 ML Estimation of Conditional Gaussian Model Parameters

The maximum likelihood (ML) estimate of the model parameters $\boldsymbol{B}$ and $\boldsymbol{b}$ from the training data is derived as a solution to a simple convex optimization problem, like ML estimation of a multivariate Gaussian. The training data $\mathcal{D} = \{(\boldsymbol{t}_1, \boldsymbol{h}_1), \ldots, (\boldsymbol{t}_N, \boldsymbol{h}_N)\}$ for a conditional Gaussian model for a given context consists of all the pairs $(\boldsymbol{t}_i, \boldsymbol{h}_i)$ of tail and head spectral feature vectors from the N unit boundaries available from the corpus for that context.

In order to facilitate the calculation, we define a $d \times (d+1)$ augmented matrix $\boldsymbol{A}$ and a $(d+1)$-dimensional vector $\boldsymbol{s}_i$, where d is the dimensionality of $\boldsymbol{t}_i$ and $\boldsymbol{h}_i$, such that,

$$\boldsymbol{A} = \left[\begin{array}{c|c} \boldsymbol{b} & \boldsymbol{B} \end{array}\right], \quad \textit{and} \quad \boldsymbol{s}_i = \left[\begin{array}{c} 1 \\ \boldsymbol{t}_i \end{array}\right], \tag{2}$$

with which we have the relationship $\boldsymbol{A}\boldsymbol{s}_i = \boldsymbol{B}\,\boldsymbol{t}_i + \boldsymbol{b}$. Thus, we obtain the estimates of $\boldsymbol{B}$ and $\boldsymbol{b}$ from the estimate of $\boldsymbol{A}$. Then the conditional Gaussian density function can be written as

$$\begin{aligned} &\mathcal{N}(\boldsymbol{h}|\boldsymbol{B}\,\boldsymbol{t} + \boldsymbol{b},\, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{h}|\boldsymbol{A}\,\boldsymbol{s},\, \boldsymbol{\Sigma}) \\ &\quad = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{h} - \boldsymbol{A}\boldsymbol{s})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{h} - \boldsymbol{A}\boldsymbol{s})\}. \end{aligned} \tag{3}$$

The log likelihood $\mathcal{L}$ with the training data $\mathcal{D}$ is, therefore,

$$\begin{aligned} &\mathcal{L}(\boldsymbol{A}, \boldsymbol{\Sigma}; \mathcal{D}) \triangleq \log \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{h}_i|\boldsymbol{A}\boldsymbol{s}_i, \boldsymbol{\Sigma}) \\ &\quad = -\frac{dN}{2}\log 2\pi - \frac{N}{2}\log|\boldsymbol{\Sigma}| \\ &\qquad -\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{h}_i - \boldsymbol{A}\boldsymbol{s}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{h}_i - \boldsymbol{A}\boldsymbol{s}_i). \end{aligned} \tag{4}$$

Taking the partial derivative of $\mathcal{L}$ with regard to $\boldsymbol{A}$, and utilizing the formula (see, e.g., [17]),

$$\frac{\partial\{(\boldsymbol{X}\boldsymbol{a} + \boldsymbol{b})^T \boldsymbol{C}(\boldsymbol{X}\boldsymbol{a} + \boldsymbol{b})\}}{\partial \boldsymbol{X}} = (\boldsymbol{C} + \boldsymbol{C}^T)(\boldsymbol{X}\boldsymbol{a} + \boldsymbol{b})\boldsymbol{a}^T,$$

we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{A}} &= -\frac{1}{2}\sum_{i=1}^{N}\{-(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1^T})(\boldsymbol{h}_i - \boldsymbol{A}\boldsymbol{s}_i)\boldsymbol{s}_i^T\} \\ &= \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}(\boldsymbol{h}_i - \boldsymbol{A}\boldsymbol{s}_i)\boldsymbol{s}_i^T. \end{aligned} \tag{5}$$

Setting the partial derivative to zero, we obtain the ML estimate of $\boldsymbol{A}$ as

$$\hat{\boldsymbol{A}} = \Big(\sum \boldsymbol{h}_i \boldsymbol{s}_i^T\Big)\Big(\sum \boldsymbol{s}_i \boldsymbol{s}_i^T\Big)^{-1}. \tag{6}$$

The covariance matrix $\boldsymbol{\Sigma}$ can be estimated as the sample covariance around the conditional mean $\hat{\boldsymbol{A}}\,\boldsymbol{s}_i$, and it reduces to

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}_i\boldsymbol{h}_i^T - \hat{\boldsymbol{A}}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{s}_i\boldsymbol{h}_i^T. \tag{7}$$

3. Robust Training with Decision-Tree Clustering

The number of the types of contexts that determines the specific conditional Gaussian (CG) model for measuring the goodness of concatenation can be very large and we often have very few training data points (or, even worse, no data points at all) available from the corpus for some types of contexts. For example, even if we assume that the context is simply determined by the phone identities of the current unit and the preceding unit, the number of possible combination is already close to 3000. In order to achieve robust training of the conditional Gaussian concatenation models, we tie the model parameters using decision tree-based context clustering. The process of parameter tying is performed by the following steps.

1. Initial CG model parameters are trained for all the distinct contexts existing in the training data.
2. CG models that have the same phone labels associated with the head spectral feature vector $\boldsymbol{h}_i$ in the boundary data pair $(\boldsymbol{t}_i, \boldsymbol{h}_i)$ are grouped and clustered using a decision tree:
   a. Parameters of all the CG models in the group are tied and associated with the root node of the decision tree.
   b. Each terminal node of the tree is examined and recursively split into two child nodes based on the context question that yields the maximum increase of the likelihood.
   The node is not split if the likelihood gain is below the pre-specified threshold or the number of training data points after split is smaller than the pre-specified minimum number of elements in the node.

Suppose we have a subset of the augmented training data $\mathcal{S} = \{(\boldsymbol{s}_1, \boldsymbol{h}_1), \ldots, (\boldsymbol{s}_n, \boldsymbol{h}_n)\}$ associated with a node, where $\boldsymbol{s}_i$ is a $(d+1)$-dimensional augmented tail vector defined in the Eq. (2). Let $\mathcal{L}_{\mathcal{S}}$ be the log likelihood, with regard to the data $\mathcal{S}$, of the model trained with $\mathcal{S}$ itself. Noting the relationship

$$\begin{aligned} &\sum_{i=1}^{n}(\boldsymbol{h}_i - \boldsymbol{A}_{\mathcal{S}}\boldsymbol{s}_i)^T \boldsymbol{\Sigma}_{\mathcal{S}}^{-1}(\boldsymbol{h}_i - \boldsymbol{A}_{\mathcal{S}}\boldsymbol{s}_i) \\ &\quad = trace(\boldsymbol{\Sigma}_{\mathcal{S}}^{-1} \cdot\, n\,\boldsymbol{\Sigma}_{\mathcal{S}}) = n \,\cdot\, d, \end{aligned}$$

where $\boldsymbol{A}_{\mathcal{S}}$ and $\boldsymbol{\Sigma}_{\mathcal{S}}$ are the augmented regression matrix and the covariance matrix trained with $\mathcal{S}$, we can reduce $\mathcal{L}_{\mathcal{S}}$ into

$$\begin{aligned} \mathcal{L}_{\mathcal{S}} &= \log \prod_{i=1}^{n} \mathcal{N}(\boldsymbol{h}_i|\boldsymbol{A}_{\mathcal{S}} \cdot \boldsymbol{s}_i, \boldsymbol{\Sigma}_{\mathcal{S}}) \\ &= -\frac{n}{2}(d\log(2\pi) + \log|\boldsymbol{\Sigma}_{\mathcal{S}}| + d). \end{aligned} \tag{8}$$

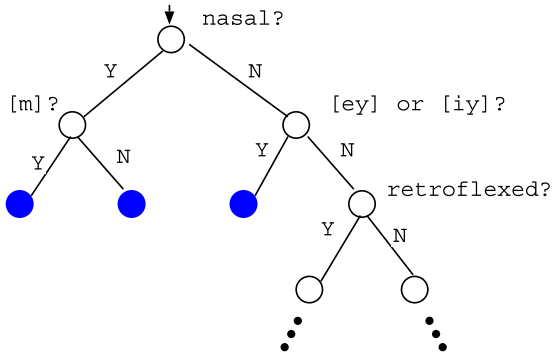


Fig. 3 A decision tree for clustering the context for the group of boundary data with phone [aa] associated with the head feature vector. Open circles represent non-terminal nodes and filled circles represent terminal nodes. Nodes are split by phonetic questions on the preceding unit.

Therefore, we see that the log likelihood with $\mathcal{S}$ depends only on the covariance matrix $\boldsymbol{\Sigma}_s$ and the number of data points n. When $\mathcal{S}$ is divided into the subsets $\mathcal{A}$ with a data points and $\mathcal{B}$ with b $(= n - a)$ data points by a context question, the increase in the log likelihood $\mathcal{G}$ becomes

$$\begin{aligned}\mathcal{G} &= \mathcal{L}_{\mathcal{A}} + \mathcal{L}_{\mathcal{B}} - \mathcal{L}_{\mathcal{S}} \\ &= \frac{1}{2}\{(a+b)\log|\boldsymbol{\Sigma}_{\mathcal{S}}| - a\log|\boldsymbol{\Sigma}_{\mathcal{A}}| - b\log|\boldsymbol{\Sigma}_{\mathcal{B}}|\}. \end{aligned} \quad (9)$$

Using the Eqs. (6) and (7), the increase $\mathcal{G}$ can be computed efficiently utilizing the sufficient statistics $\sum_i \boldsymbol{h}_i \boldsymbol{s}_i^T$, $\sum_i \boldsymbol{s}_i \boldsymbol{s}_i^T$, $\sum_i \boldsymbol{h}_i \boldsymbol{h}_i^T$, and $\sum_i \boldsymbol{s}_i \boldsymbol{h}_i^T$. We compute these sufficient statistics for all the un-tied models in the step 1 of the decision tree-based clustering process described earlier. The likelihood at any node can be computed by reusing these sufficient statistics without direct reference to the training data points.

Figure 3 depicts part of the decision tree grown for clustering the context for the group of boundary data in which the head feature vectors have phone label [aa], obtained through the training of CG models in the experiment presented in Sect. 5.

4. Relationships with Other Approaches

The proposed measure for the cost of concatenation between the *tail* feature vector $\boldsymbol{t}$ from the preceding unit and the *head* feature vector $\boldsymbol{h}$ from the current unit is the negative of the log probability given by the conditional Gaussian model expressed as

$$\begin{aligned}&\log \mathcal{N}(\boldsymbol{h}|\boldsymbol{B}\,\boldsymbol{t} + \boldsymbol{b},\ \boldsymbol{\Sigma}) \\ &\quad = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| \\ &\qquad -\frac{1}{2}(\boldsymbol{h} - (\boldsymbol{B}\,\boldsymbol{t} + \boldsymbol{b}))^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{h} - (\boldsymbol{B}\,\boldsymbol{t} + \boldsymbol{b})), \end{aligned} \quad (10)$$

where the model parameters $\boldsymbol{B}$, $\boldsymbol{b}$, and $\boldsymbol{\Sigma}$ depend on the context (in the current experiment, phone identity of the current unit and the preceding phone context). The integer d is the dimensionality of the vectors $\boldsymbol{h}$ and $\boldsymbol{t}$. By comparing this equation with the formulas of other distance measures, we show that the proposed method has interesting relationships with other approaches.

4.1 Squared Euclidean Distance

Euclidean distance is a widely used distance measure and its square is sometimes preferred over the standard Euclidean distance when faster computation is required. If we set the transformation matrix $\boldsymbol{B}$ to the identity matrix (I), the constant $\boldsymbol{b}$ to zero vector, the covariance matrix $\boldsymbol{\Sigma}$ also to the identity matrix and neglect the constant terms, we note that the negative of the score given by Eq. (10) reduces to the square of the Euclidean distance between $\boldsymbol{h}$ and $\boldsymbol{t}$ that can be expressed as

$$D^2_{euc} = (\boldsymbol{h} - \boldsymbol{t})^T(\boldsymbol{h} - \boldsymbol{t}). \quad (11)$$

4.2 Donovan's Approach

In [11], Donovan proposed a distance measure between the vector e at the end (i.e. tail) of one segment and the vector s at the start (i.e. head) of the next segment. For this purpose, he clustered the pairs of frames across the boundaries using a decision tree by asking broad class questions about the preceding and following phonetic identity and the location of the boundary within the phone, and calculated the mean and the covariance matrix within each leaf of the tree. He describes it "a decision-tree-based context-dependent Mahalanobis distance", which is expressed as

$$D^2 = \sum_{i=1}^{d}\left[\frac{e_i - s_i - \mu_i^l}{\sigma_i^l}\right]^2, \quad (12)$$

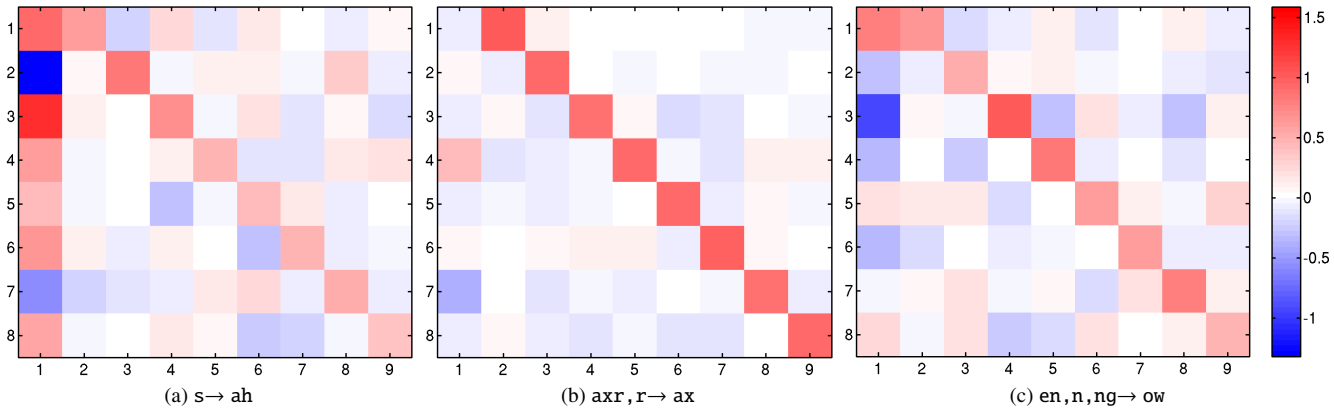
where d is the dimensionality of the data, μ_i^l is the i-th element of the mean vector in leaf l, σ_i^l is the i-th diagonal element of the covariance matrix for leaf l.

Looking at the Eqs. (10) and (12), we note that (10) becomes equivalent to (12) if we set $\boldsymbol{B}$ to the identity matrix and neglect the second term with the determinant of the covariance matrix, also assuming that the elements of the feature vectors are independent to each other. In other words, Donovan's distance measure is similar to the conditional Gaussian-based concatenation model with the conditional mean formed by just the addition of the constant $\boldsymbol{b}$ and no transform by the matrix $\boldsymbol{B}$.

To wrap up, we see that the proposed method is a generalization of the approaches described above, endowed with a stronger descriptive power from the equations in Sects. 4.1 and 4.2. We will also look at the effectiveness of the proposed method quantitatively in the next section.

5. Experiments

We trained the conditional Gaussian concatenation models using the speaker SLT database of the CMU Arctic speech



(a) s→ ah (b) axr,r→ ax (c) en,n,ng→ ow

Fig. 4 Graphical representations of the $8 \times (1 + 8)$ augmented regression matrices $\boldsymbol{A} = [\boldsymbol{b}|\boldsymbol{B}]$ trained using the Arctic SLT corpus for concatenation boundaries of (a) from [s] to [ah], (b) from [axr]/[r] to [ax], and (c) from [en]/[ng]/[n] to [ow]. Small squares represent matrix elements and the color bar on the right shows the mapping from the element's value to its color. Darker squares have larger absolute values. Red means positive and blue means negative.

databases [18]. It is spoken by a female speaker of American English and consists of 1,132 utterances. The total duration is roughly 50 minutes. The phone inventory we used is a detailed phone set consisting of 53 phones that includes the phone set used in the CMU pronouncing dictionary [19]. All the units to be selected and concatenated are phone-sized in this experiment. The decision tree-based clustering of the context for tying model parameters was performed for each of the groups of boundary data that has the same phone label associated with the head feature vector and the phone identity of the previous unit was used as the context to be clustered. The head and the tail feature vectors were the spectral features averaged over a 10 ms interval (two 5-ms frames) at the both ends of the unit. As a spectral feature vector, we used 14 MFCC coefficients with dimensionality reduced to 8 by principal component analysis. For the stopping criteria of node divisions in the decision tree-based clustering, the likelihood gain threshold was set to 1.0 empirically through experiments and the minimum number of data points was set to 17, which is $2 \times d + 1$ with d (= 8) being the dimensionality of the head and tail feature vectors described above, to avoid rank deficiency of the covariance matrix. As a result, the whole 2,809 (= 53^2) combinations of the tail and head phone labels were clustered into 677 clusters. Consequently, 677 conditional Gaussian models were trained.

Figure 4 depicts three examples of the augmented regression matrices of the conditional Gaussian models trained. In the left matrix (Fig. 4 (a)), which comes from the phonetic context [s] for [ah], a typical context involving abrupt change in the spectral shape, we see that the constant vector part $\boldsymbol{b}$ is dominant in the linear transform $\boldsymbol{Bt} + \boldsymbol{b}$, whereas we also note a slight diagonal pattern in the regression matrix. On the other hand, the diagonal components of the regression matrix $\boldsymbol{B}$ are dominant in the transition from [axr] or [r] to [ax] (Fig. 4 (b)), suggesting that the spectral shape is very similar on the both sides of the boundary. In Fig. 4 (c), we notice significant contributions from both of the constant vector $\boldsymbol{b}$ and the regression matrix $\boldsymbol{B}$ for the boundary of a nasal consonant ([en], [n], or [ng]) and the vowel [ow]. These results are in concordance with our discussion in Sect. 2 and we can expect that a reasonable measure for scoring the goodness of concatenation is achieved which gives high scores to candidate units with head feature vectors close to the conditional mean predicted by the model rather than always preferring those units with the head feature vector similar to the tail feature vector of the preceding unit.

In order to investigate the effectiveness of the proposed approach to concatenation modeling, we performed several kinds of evaluation experiments using Euclidean distance as the baseline for comparison, which has been reported to be a good predictor of perceived discontinuity when measured on Mel-cepstral feature parameters [10]. For synthesizing the utterances, we made use of the speech synthesizer reported in [14] trained also with the Arctic SLT database. In this synthesizer, the total cost C is the sum of three kinds of target costs (c_d^t for duration, c_f^t for F_0, and c_s^t for spectrum) and the spectral concatenation costs c_s^c,

$$C = \sum_{k=1}^{N} \{c_d^t(u_k) + c_f^t(u_k) + c_s^t(u_k)\} + \sum_{k=2}^{N} c_s^c(u_{k-1}, u_k), \tag{13}$$

where the concatenation cost c_s^c with the proposed method is defined as

$$c_s^c(u_{k-1}, u_k) = -w \cdot \log P(\boldsymbol{h}(u_k)|\boldsymbol{t}(u_{k-1}), c_k), \tag{14}$$

and when the Euclidean distance is used, it is defined to be

$$c_s^c(u_{k-1}, u_k) = w \cdot \|\boldsymbol{h}(u_k) - \boldsymbol{t}(u_{k-1})\|, \tag{15}$$

where w is a weighting coefficient.

5.1 Objective Evaluation Experiments

We first investigated the effectiveness of the proposed

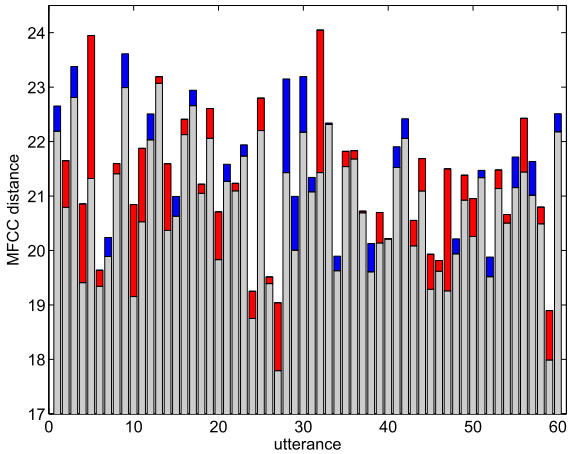


Fig. 5 MFCC distances between synthetic and natural speech for the baseline (euc) and the proposed approach (cg) for 60 open sentences. The vertical axis represents the distance calculated using the Eq. (16) with dynamic time warping. 60 points along the horizontal axis represent the 60 utterances. When the distance to natural speech is larger with euc, the upper part of the bar is in red and the lower part is in gray. In this case, the whole bar (red + gray) represents the distance with euc and the just the gray part represents the distance with cg. Therefore, the red part means the difference (euc − cg). When the distance is larger with cg, the upper part of the bar is in blue and the same applies for the interpretation of the bar. In this case, the blue part means the difference (cg − euc).

method through an objective evaluation in which we compared the closeness of the synthetic speech to natural speech as measured by the distance between the MFCC parameter sequences extracted from them. Fourteen MFCC coefficients excluding the 0-th coefficient were extracted from each speech waveform with the frame rate of 5 ms. The frame-wise distance $d(\boldsymbol{x}, \boldsymbol{y})$ of the two frame vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ was calculated as the Euclidean distance between these two 14-dimensional vectors:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{14} (x_j - y_j)^2 \right)^{1/2}, \tag{16}$$

where x_j and y_j are the j-th components of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. The differences of the lengths of the parameter sequences were absorbed using dynamic time warping [20]. In the objective evaluation, we compared the effectiveness of the proposed method with the related approaches presented in the last section as well as with the baseline. The weight on the concatenation cost has been optimized with regard to MFCC distance for each method independently using 40 separate open sentences. Then, 60 open sentences were synthesized for the objective evaluation which consisted of 30 conversational sentences and 30 sentences extracted from novels, that were used in the Blizzard Challenge 2005 [21]. Figure 5 plots the utterance-wise average MFCC distances between the synthetic and natural speech for the baseline and the proposed approach with the 60 open sentences, highlighting the difference of distances in red (when the distance is larger with the baseline) and blue (when the distance is larger with the proposed method). Table 1 shows the means and the standard deviations (s.d.) of the MFCC distances for

Table 1 Objective evaluation results using 60 open sentences. Mean and standard deviation (s.d.) of the utterance-wise MFCC distances between natural and synthesized speech are presented with various methods. cg stands for the proposed method using conditional Gaussians, whereas cg (diag) is the same method but the matrix $\boldsymbol{B}$ is restricted to be diagonal. donovan represents Donovan's approach. sq euc stands for the square of the Euclidean distance, and euc stands for the Euclidean distance.

	mean	s.d.
cg	21.00	1.32
cg(diag)	21.11	1.23
donovan	21.16	1.17
sq euc	21.19	1.17
euc	21.24	1.18

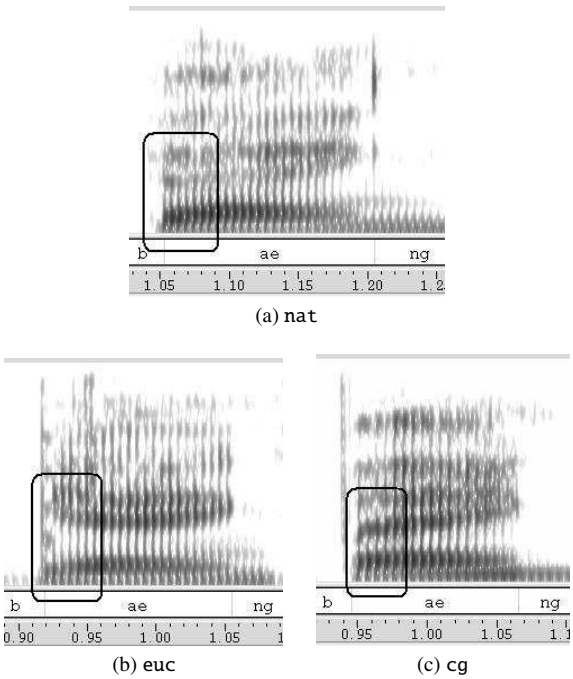
the proposed and various related approaches introduced in the last section as well as the baseline. As we see from Fig. 5 and Table 1, the proposed method (cg) achieves a smaller distance to natural speech than the baseline (euc) and it was statistically significant (p-value = 0.014). We also see that the proposed approach achieves synthesized speech closer to natural speech compared to the related methods that can be interpreted as special degenerate cases of the proposed method. We also note that the non-diagonal elements of the matrix $\boldsymbol{B}$ in (10) representing the dependencies among feature dimensions is effective when we compare the entries cg and cg(diag). The differences between the proposed method (cg) and other methods (i.e. donovan, sq euc, and euc) were all statistically significant at the 5% level.

5.2 Listening Tests

In order to evaluate the perceptual performance of the proposed method, we performed a listening test. For the proposed method, the weight for the concatenation cost w was set to be 1.0 based on rough informal listening. To be fair with the baseline, we optimized the weight w for the baseline by a more elaborate experiment. Specifically, we preliminarily synthesized ten utterances with varying values of the weight w and picked the one that yielded the best sounding synthetic speech by informal listening. We first tested a few values of w preliminarily and noted that the speech quality is quite insensitive to the changes on the linear scale. Therefore, we first varied the weights different from each other on an exponential scale, namely, 0.1, 1.0, 10.0, 100.0, and found out that 1.0 was the best among them. Then we tested several values around 1.0, which were 0.4, 0.7, 1.0, 2.0, 3.0, 4.0. It turned out that 1.0 was the best among them, which was subtly better than 0.7 and 2.0. A set of twenty open sentences (10 conversational sentences and 10 sentences from novels) extracted from the Blizzard Challenge 2005 test set [21] were used in the listening test. The eight listeners were students and researchers at a university in the United States who use English in their daily life. They consist of both native and non-native speakers of American English. These eight subjects listened to the speech synthesis output from two synthesizers, one of which adopting the baseline and the other with the proposed models for concatenation cost. They were asked to judge how good the speech

Table 2 5-level mean opinion scores with standard deviations (s.d.) for the two concatenation modeling approaches, (1) the baseline with Euclidean distance (euc), and (2) the proposed method with conditional Gaussian-based concatenation models (cg).

	mean	s.d.
euc	2.44	0.837
cg	2.97	0.893



(a) nat

(b) euc (c) cg

Fig. 6 Spectrogram fragments corresponding to the first syllable ([b ae ng]) of the word "Bangkok" from the natural utterance ((a) nat), synthesized speech by baseline ((b) euc), and the proposed method ((c) cg). The whole sentence was ''I'd like to fly to Bangkok.''.

sounded and assigned a score of 1 to 5 to each utterance. Linguistic expressions such as "good" and "very good" were not associated with the scores in the directions given to the listeners. The results of the listening test is summarized in Table 2. The mean opinion score with the proposed method turned out to be significantly higher than the baseline at the 1% level by the paired t-test, with a p-value of 5.18×10^{-11}.

Compared with the objective evaluation results presented in Sect. 5.1, the proposed method is conspicuously better than the baseline in the listening test. Therefore, we closely examined some of the utterances with which the speech synthesized with the proposed method has a significantly better mean opinion score in the listening test whereas the objective measure is not as good. An example of this is depicted in Fig. 6. In this example, MOS is 3.25 with the baseline ((b) euc) and 3.75 with the proposed method ((c) cg) and we notice a more natural formant transition from [b] to [ae] with the proposed method which appears more similar to the natural speech when we compare the regions surrounded by round-cornered rectangles in the figure. In the MFCC distance, however, this local advantage seems to be absorbed in the accumulation of local distances

Table 3 Objective evaluation results using 60 open sentences with mean and standard deviation (s.d.) of the utterance-wise MFCC distances for the open test set with speaker BDL. 'euc' represents the baseline with Euclidean distance and 'cg' represents the proposed method that employs conditional Gaussian-based concatenation models.

	mean	s.d.
euc	22.85	1.55
cg	22.27	1.29

over the utterance and the average distance is slightly better (21.29) for the baseline compared to the proposed method (22.13) with this example.

5.3 Effectiveness Across Speakers

In order to confirm the effectiveness of the proposed method across speakers, we developed another voice and trained the conditional Gaussian concatenation models using speaker BDL of the CMU Arctic speech databases [18]. It is spoken by a male speaker of American English and consists of the same sentences as the SLT database. The models are trained in the same way as SLT and the objective evaluation was conducted in the same way as described in Sect. 5.1. Table 3 presents the results of the objective evaluation with 60 open sentences. As seen in the table, proposed method achieved a smaller distance to natural speech with this speaker as well and it was statistically significant (p-value = 2.09×10^{-5}).

6. Discussion

In recent years, hybrid approaches of unit selection and HMM-based speech synthesis have been emerging and shown to be effective [22]–[24]. They appear to share the background motivation with the current paper in the use of acoustic models for target and concatenation modeling rather than relying on heuristic knowledge. The original contributions of the current paper that differentiate it from other works include the explicit use of the dependencies among feature dimensions in the matrix $\boldsymbol{B}$ which is used to obtain the conditional mean, and the clustering method for context tying that directly uses the objective of maximum likelihood of the concatenation model alone.

7. Conclusion

In this paper, we presented a novel probabilistic approach to concatenation modeling using conditional Gaussian models. We presented a maximum likelihood estimation formula for the models and a robust and efficient training scheme using decision-tree based context clustering. We implemented the proposed method with the CMU Arctic speech databases and confirmed the effectiveness of the proposed method through an objective evaluation in which the closeness of the synthetic speech to natural speech was measured as well as a subjective listening test. We also presented the relationships of the proposed method with other approaches and showed that the proposed method has a flexible modeling power and comprises various other concatenation cost

approaches as special cases of it. In the current paper, the effectiveness of the proposed method is confirmed using the framework of Blizzard Challenge 2005, in which the corpus size is around one hour. Examination of the effectiveness of the proposed approach with larger corpora, such as a ten-hour speech corpus, will be one of the important future directions to take.

Acknowledgements

The authors would like to thank people at MIT Computer Science and Artificial Intelligence Laboratory who participated in the listening test. The authors are also grateful to Prof. Alan Black at Carnegie Mellon University for letting us use a part of test data for Blizzard Challenge 2005 and to Prof. Keiichi Tokuda and Dr. Tomoki Toda for insightful discussions.

References

[1] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," Proc. ICASSP 1988, pp.449–452, 1988.

[2] N. Iwahashi, N. Kaiki, and Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," IEICE Trans. Fundamentals, vol.E76-A, no.11, pp.1942–1948, Nov. 1993.

[3] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. ICASSP '96, pp.373–376, 1996.

[4] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft Mulan – a bilingual TTS system," Proc. ICASSP 2003, pp.I–264–I–267, 2003.

[5] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "Ximera: A new TTS from ATR based on corpus-based technologies," Proc. ISCA 5th Speech Synthesis Workshop, pp.179–184, 2004.

[6] Z.J. Yan, Y. Qian, and F.K. Soong, "Rich-context unit selection (RUS) approach to high quality TTS," Proc. ICASSP 2010.

[7] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis," Proc. ICASSP 2004, pp.657–660, Montreal, Canada, May 2004.

[8] J. Wouters and M. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," Proc. ICSLP 98, pp.2747–2750, Sydney, Australia, 1998.

[9] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," IEEE Trans. Speech Audio Process., vol.9, no.1, pp.39–51, 2001.

[10] Y. Stylianou and A.K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," Proc. ICASSP 2001, Salt Lake City, USA, 2001.

[11] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," Proc. 4th ESCA Tutorial and Research Workshop on Speech Synthesis, Scotland, Sept. 2001.

[12] J. Vepa and S. King, "Join cost for unit selection speech synthesis," in Text to Speech Synthesis, ed. A. Alwan and S. Narayanan, Prentice Hall, 2004.

[13] J. Bilmes, "Graphical models and automatic speech recognition," in Mathematical Foundations of Speech and Language Processing, Springer-Verlag, 2003.

[14] S. Sakai and H. Shu, "A probabilistic approach to unit selection for corpus-based speech synthesis," Proc. Interspeech 2005, pp.81–84, Lisbon, Portugal, Sept. 2005.

[15] S. Sakai, "Additive modeling of english F0 contour for speech synthesis," Proc. ICASSP 2005, pp.I-277–I-280, Philadelphia, PA, March 2005.

[16] S. Sakai, "Fundamental frequency modeling for speech synthesis based on a statistical learning technique," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.489–495, March 2005.

[17] K.B. Petersen and M.S. Pedersen, The Matrix Cookbook, Technical University of Denmark, Oct. 2008. "http://www2.imm.dtu.dk/pubdb/p.php?3274".

[18] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. CMULTI-03-177, Language Technologies Institute, CMU, 2003.

[19] "The CMU pronouncing dictionary." Web site: http://www.speech.cs.cmu.edu/cgi-bin/cmudict, Papers not available as of writing.

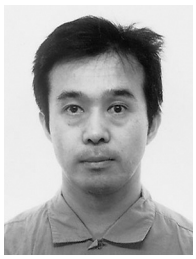
[20] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, April 1993.

[21] A. Black and K. Tokuda, "Blizzard challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," Proc. Interspeech 2005, pp.77–80, Lisbon, Portugal, 2005.

[22] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe, "Recent improvements on microsoft's trainable text-to-speech system - whistler," Proc. ICASSP 1997, pp.959–962, 1997.

[23] Z.H. Ling and R.H. Wang, "HMM-based unit selection using frame sized speech segments," Proc. Interspeech 2006, Pittsburgh, pp.2034–2037, Sept. 2006.

[24] Z.H. Ling and R.H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," Proc. ICASSP-07, pp.1245–1248, Honolulu, April 2007.



Shinsuke Sakai received his B.E. and M.E. degrees from Kyoto University in 1982 and 1984, respectively. He joined NEC Corp. in 1984, where he worked on various aspects of speech and natural language processing for machine translation and speech recognition. Between 1991-1993, he was a visiting scientist at Massachusetts Institute of Technology, Cambridge, MA., U.S.A. Between 2002-2004, he was with the Laboratory for Computer Science and Computer Science and Artificial Intelligence Laboratory, where he worked on speech synthesis. He has been affiliated with the graduate school for informatics at Kyoto University from 2005 to 2008. He worked as a senior researcher at ATR Spoken Language Communication Laboratories in Kyoto, Japan between 2006-2009 and as an expert researcher at Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology between 2009-2011. Currently, he is affiliated with Academic Center for Computing and Media Studies, Kyoto University. His research interests include exploring speech technology that can be used in various daily life situations, such as in the office, living room and on the street. He is a member of the Information Processing Society of Japan, the Acoustical Society of Japan, and the IEEE.

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Affiliated Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR, currently National Institute of Information and Communications Technology. (NICT). He has published more than 200 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been managing several speech-related projects in Japan including a free large vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/). Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. From 2003 to 2006, he was a member of the IEEE SPS Speech Technical Committee. He was a general chair of the IEEE Automatic Speech Recognition & Understanding workshop (ASRU-2007). He is a senior member of IEEE.

Hisashi Kawai received B.E., M.E., and D.E. degrees in electronic engineering from The University of Tokyo, in 1984, 1986, and 1989, respectively. He joined Kokusai Denshin Denwa Co. Ltd. in 1989. He worked for ATR Spoken Language Translation Research Laboratories from 2000 to 2004, where he directed the development of a corpus-based text-to-speech synthesis system named XIMERA as the head of the department of speech synthesis. Since October 2004, he worked for KDDI R&D Laboratories Inc., where he was engaged in the management of research and development of speech information processing, speech quality control for telephone, speech signal processing, and acoustic processing. Since 2009, he has been working for National Institute of Information and Communications Technology (NICT). He is a member of the Acoustical Society of Japan (ASJ) and IEEE.