Substitution of state distributions to reproduce natural prosody on HMM-based speech synthesizers

Nobuyuki Nishizawa, Tsuneo Kato
KDDI R&D Laboratories Inc., Japan
no-nishizawa@kddilabs.jp, tkato@kddilabs.jp

Abstract
An extension of HMM-based speech synthesis is proposed to reproduce natural speech sounds. For compression of large amounts of speech, the use of speech synthesizers has an advantage in terms of the size of compressed data. However, the quality of synthetic speech is often inferior to that of speech compressed by general-purpose speech codecs such as CELP, where prosodic features are reproduced more accurately. Therefore, we propose adding complementary information to reproduce natural prosody. In the proposed method, inappropriate state feature vectors of HMMs determined by the conventional speech synthesis method are substituted by other vectors bound to the decision trees. The experimental results indicated that substitution of 20% of state feature vectors reduces root mean squared error (RMSE) in log F0 to 0.3 semitones, which is approximately 15% of RMSE without substitution.

Index Terms: HMM-based speech synthesis, vector substitution, speech data compression

1. Introduction
Several mobile phones or digital home appliances, for example, digital video recorders, provide voice guidance. Although prerecorded speech sounds are now compressed by general-purpose speech codecs such as CELP, where prosodic features are reproduced more accurately. Therefore, we propose adding complementary information to reproduce natural prosody. In the proposed method, inappropriate state feature vectors of HMMs determined by the conventional speech synthesis method are substituted by other vectors bound to the decision trees. The experimental results indicated that substitution of 20% of state feature vectors reduces root mean squared error (RMSE) in log F0 to 0.3 semitones, which is approximately 15% of RMSE without substitution.

Among several speech synthesis methods, hidden Markov model (HMM)-based speech synthesizers[2][3] are suitable for such purpose. This is because the footprint size of the system is small enough for mobile devices, and the computational cost for speech synthesis is comparable to CELP decoders. The quality of generated sounds seems sufficient for practical use[4].

However, in subjective comparison between the output sounds of speech codecs and speech synthesizers, the latter is often inferior. This may be because prosodic features are reproduced more accurately in the output of speech codecs. Unnaturalness in prosodic features often causes overall bad impressions of synthetic sounds.

Therefore, we propose an extension of general-purpose HMM-based speech synthesizers to reproduce natural speech sounds using complementary information that was extracted from the natural speech sounds of the target. In the proposed method, while most parts of speech are generated by the conventional speech synthesis method from speech synthesis symbols, and the indices only of inappropriate parts are overwritten by complementary information. By the proposed method, feature vectors for F0 can be coded more effectively in size.

The rest of the paper is organized as follows. First, in Section 2, applications in which the proposed method is necessary are introduced. Section 3 explains algorithms of parameter generation for HMM-based speech synthesis as the basis of the proposed method. Section 4 describes algorithms of the proposed method, and Section 5 illustrates our speech synthesis system with the proposed method. Section 6 gives objective and subjective evaluations of the proposed method. Finally, section 7 concludes the paper.

2. Expected applications
For speech message sounds from mobile devices, text-to-speech (TTS) systems are required only for a few services, for example, reading E-mail or WWW contents. For most services, prerecorded speech sounds can be used. For example, notifications for telephone calls or E-mail by speech sounds are often demanded especially for business use, where notifications solely for categories of events are required to protect secret or private information. In another scenario, GUI operations can be guided by speech outputs. Most mobile devices cannot present both the GUI itself and detailed text messages for directions of the GUI simultaneously due to the narrowness of displays. Also in such cases, prerecorded sounds, which might be generated by TTS systems to reduce costs for speech recording, are applicable. However, the size of software and data that can be additionally installed to mobile devices is often extremely restricted, for example, to several hundreds of kilobytes. Consequently, more efficient speech data compression than speech codec such as CELP is often required.

For another application of high-efficient speech data compression, speech data is expected to be embedded into printed papers via two-dimensional bar code. In Japan, QR code, which
is a two-dimensional bar code standardized at JIS (Japanese Industrial Standards)[7] and ISO[8], is commonly used, and most mobile phones equip cameras and QR-code decoding software. Figure 1 shows examples of QR code. QR codes into which URLs for WWW sites are encoded are often printed in advertisements. Similarly, compressed speech data may be embedded into QR codes. However, the efficiency of a codec such as CELP is insufficient for such purpose. Although the maximum size of data that can be encoded into a QR-code is 23,648 bits by 57x57 cells, such codes are so complicated that the image is difficult to capture. For most mobile phones, codes for 2,192 bits by 57x57 cells are officially supported, where only a few seconds of speech sounds can be encoded by speech codecs such as CELP.

The major distinction of our method is that the conventional speech synthesizers from symbolized phonemic and prosodic information are used as the basis of speech data modeling. Therefore, frameworks for the proposed speech data compression can also be used as those for speech synthesis. Even when target HMMs for the proposed method are not installed in a device, symbolized phonemic and prosodic information encoded into compressed data can be used for the conventional speech synthesizers on the device. This sense will improve the usability of services. For example, some users may use HMMs that generate preferred voices ignoring complementary information. In some cases, to encode a large amount of information for speech synthesis into a small storage, size of complementary information may be reduced. In extreme cases, no complementary information can be encoded. In other cases, complementary information for two or more sets of HMMs may be concurrently encoded into a store for a speech message where symbolized phonemic and prosodic information is shared among the sets of HMMs. In this configuration, users can select HMMs to decode speech sounds.

In this study, although reproduction of natural F0 contours is discussed, synthetic F0 contours may be used for some applications. For example, F0 contours generated from more accurate HMMs that are not installed in the target devices can be used for the target contours to improve the naturalness of prosody. In a possible scenario, the proposed method can also be used instead of updating HMMs that have been pre-installed in the devices and generate inappropriate prosody.

3. HMM-based speech synthesis

In our HMM-based speech synthesizer, speech waveforms are produced by mel-cepstral coefficients, voiced/unvoiced information and log F0, similar to [3]. Speech waveforms are synthesized with an MLSA (Mel log spectrum approximation) filter[9] excited by impulse trains for voiced sounds or white noise for unvoiced sounds. Parameters for the filter and sources are sampled every 5 milliseconds and modeled by HMMs.

An HMM for an utterance is built by concatenating phone HMMs. In this study, the structure of each phone HMMs is 5-state left-to-right with no skip. In the speech synthesizer, distributions of the numbers of repeated frames for each state of HMM are stored instead of state transition probabilities.

In the following sections, algorithms to generate parameters for speech synthesis are explained.

3.1. Parameter generation

c_{t}$ and $o_{t}$ denote static feature vectors and observation vectors for HMMs at time $t$, respectively. For speech synthesis, the sequence of static feature vectors $C = [c_{1}, c_{2}, ..., c_{T}]$ is required. To generate smooth sequences, observation vector $o_{t}$ for speech synthesis often includes not only $c_{t}$, but also dynamic feature vectors of $c_{t}$. In this study, $o_{t}$ consists of the static feature vector $c_{t}$ and two dynamic feature vectors $\Delta c_{t}$ and $\Delta^{2}c_{t}:

\begin{equation}
\begin{align*}
\alpha_{t}^{T} &= [c_{t}, \Delta c_{t}, \Delta^{2}c_{t}] \\
\delta_{t} &= -1/2 c_{t-1} + 1/2 c_{t+1} \\
\Delta c_{t} &= c_{t-1} - c_{t} + c_{t+1}
\end{align*}
\end{equation}

Distributions of $o_{t}$ are modeled at each state of HMMs. $C$ for speech synthesis is determined by

\begin{equation}
C = \arg \max_{C} P(O|q, \lambda)
\end{equation}

where

\begin{equation}
O = [o_{1}, o_{2}, ..., o_{T}], \quad \lambda \quad \text{and} \quad q \quad \text{are HMM} \quad \text{and} \quad \text{the state transition of the HMM, respectively. Similar to [10], sub-optimal} \quad q \quad \text{estimated only from the distributions of the state durations of the HMM is used for the estimation for} \quad C \quad \text{in this study.}
\end{equation}

3.2. Determination of distributions of observation vectors

In our speech synthesizer, the state transition of HMMs is dealt with deterministically, not probabilistically, like many HMM-based speech synthesizers for the sake of simplicity. Consequently, distributions of the observation vector at a frame are exactly equal to the distributions bound to the state of the phone HMM corresponding to the frame.

In this study, a single-mixture Gaussian distribution is assumed for $o_{t}$ for each target parameter. Therefore, the distribution of $o_{t}$ for frame $t$ is represented by a state feature vector $v_{q}$ that is composed of elements of the mean vector and the covariance matrix bound to the corresponding state $q$ of the HMM.

In speech synthesis, phones should be classified with respect to both phonetic and prosodic environment for high-quality synthetic sounds. Since modeling for all possible phones with such detailed classifications is impractical, a mechanism to predict HMMs for any phone is indispensable. Therefore, tree-based state tying derived from context-oriented clustering is adopted for the phone HMMs. In this study, state durations for phones are also modeled using a similar method, where a set of parameters of state duration distributions for a phone is represented by a vector.

Thus, state transition $q$ is determined by the decision trees from the target phone with respect to the phonetic and prosodic context. In our speech synthesizer, components of $v_{q}$ for mel-cepstral coefficients, and log F0 including voiced/unvoiced information are independently determined by independent decision trees, i.e. $v_{q}$ is split into two vectors practically. Therefore, since each phone HMM is composed of five states and one
vector represents state duration distributions for a phone HMM, all features for a phone are represented by 11 vectors that are bound to the leaf nodes of the independent decision trees.

4. Substitution of state feature vectors to reproduce the target sounds

To synthesize the desired speech sounds on HMM-based speech synthesis frameworks, feature vector sequence \( C \) determined by the algorithms for the conventional HMM-based speech synthesis is modified by the following methods: (a) direct modification of elements of \( C \), (b) modification of distribution parameters for \( \alpha_t \) at frame \( t \), and (c) modification of distribution parameters for \( \beta_q \) bound to state \( q \). Among the three methods, (c) is the most difficult for accurate modification, but can be the most effective in the size. In the following, (c) is examined to reduce the size.

4.1. Overview of the proposed method

If all modifications are represented as the replacement of state feature vectors by other vectors included in HMMs, coding only for the replaced states and new vector indices is necessary without coding for each element of vectors where no additional codebooks for modification are necessary. Since this configuration can be easily implemented as an extension of the conventional HMM-based speech synthesizers, we adopt this configuration in this study.

Figure 2 illustrates our proposed method with the substitution of state feature vectors schematically. In the proposed method, state feature vectors for speech synthesis are basically determined from the target speech synthesis symbols used for the conventional speech synthesis, and only inappropriate parts of speech are additionally modified. In practice, feature vector indices for each of the phone HMMs are overwritten by complementary information for modification. In encoding speech sounds, these complementary information as vector indices are greedily sought in this study. Details of the search algorithm are introduced in the next section.

In the search, two criteria must be adoptable. One is based on maximization of the likelihood function \( P(O|q, \lambda) \). The other is based on the minimization of the distance between the target \( C \) extracted from the target sound, and estimated \( C \) by HMM-based speech synthesis techniques. Search based on the former is much easier than that based on the latter since the latter criterion requires re-computation of \( C \) at every experimental substitution of vectors. However, the former is expected to be unsuitable for our system. This is because the observation vectors consist of not only static features but also dynamic features. State feature vectors that have dynamic features with relatively small variances are likely to make global error worse due to reproduction only of dynamic features in parameter generation, even while such state feature vectors can be optimal in \( P(O|q, \lambda) \). In preliminary evaluation for the used HMM-based speech synthesis system, extremely small variances of dynamic parameters were observed in state feature vectors especially for F0. Since generated F0 contours depend strongly on the dynamic feature of such vectors rather than the static feature, total error between the generated contour and the target is often deteriorated.

Consequently, error between the target and the generated parameters is evaluated for the search in this study.

![Figure 2: Substitution of state feature vectors in the proposed method.](image)

4.2. Algorithm for speech coding

As one of the simplest search algorithms, a coding algorithm based on the greedy search algorithm is introduced in this study. The procedure is as follows:

**STEP 1:** mel-cepstral coefficient and F0 are extracted from the target speech sounds, and their dynamic features \( \Delta \) and \( \Delta^2 \) are calculated. Speech synthesis symbols for the target speech are also extracted from speech sounds. In this study, for accuracy, all symbols were manually dictated.

**STEP 2:** An utterance HMM corresponding to the speech synthesis symbols is built by concatenating phone HMMs. Sequences of the vectors extracted from speech sounds are temporally aligned to states of the utterance HMM by Viterbi algorithm. For each state, the number of frames for the target sounds (i.e., state duration) is explicitly encoded into complementary information to reproduce the state duration.

**STEP 3:** For all states of the utterance HMM, all possible substitutions of state feature vectors are experimentally examined to search the optimal vector substitution. For each experimental vector substitution, the feature vector sequence \( C \) is computed under the restriction of state durations determined in **STEP 2**. After all experimental examinations, the best substitution for the utterance is encoded as complementary information. In this study, frame-wise root mean squared errors (RMSE) of log F0 are evaluated for the substitution. The corresponding state feature vector of the utterance HMM is also updated. If no better substitution is found, the procedure ends.

**STEP 4:** When the number of substituted vectors exceeds the limit number, the procedure is terminated. Otherwise, it is back to **STEP 3**.

To measure error between the generated contour and the target for each frame, these contours must be temporally aligned. In **STEP 2**, state durations of the utterance HMM are forcibly determined by ML-based Viterbi alignment with target natural speech sounds in which not only F0 but also mel-cepstrum are evaluated. In the forced-alignment, state transition probabilities rather than distributions of state durations are taken into account for simplicity.
Since substitutions by all vectors bound to all states of phone HMMs are experimentally examined for every state of the target utterance HMM in STEP3, huge computational cost is required in the above algorithm. For example, the computational time to generate the F0 contour illustrated in Figure 3 in Section 6 exceeds a thousand times the real time even when a PC with a 3-GHz CPU is used. However, the computational cost is not a considerable problem for the applications illustrated in Section 2. Reduction of computational costs will be discussed in a future study.

5. Development of speech synthesis system with the proposed method

We developed an HMM-based speech synthesis system with the proposed vector substitution mechanism. While the proposed method is applicable to the improvement of not only F0, but also other features, for example, spectral features or durations, reproduction of the original F0 contour was focused on as the first step of the proposed method.

As speech synthesis symbols for the system, the kana-level notation defined in JEITA (Japan Electronics and Information Technology Industries Association) IT-4002 Speech synthesis symbols[11] was adopted. In the kana-level notation, Japanese kana characters with a devoicing symbol, and prosodic boundary and accent nucleus symbols are used to describe both phonemic and prosodic information simultaneously. For a more efficient coding system, we experimentally designed a packed expression that encodes each of the phonemes for one mora, a devoiced mark, a prosodic boundary symbol, and an accent nucleus symbol into one 8-bit value. When pronunciations of 53 sentences of ATR503[12] spoken by a female narrator are encoded, speech synthesis symbols can be packed to 11.5 bits per mora.

On the other hand, the coding of vector indices is simpler. In our developed system, whether the vector is substituted or not is encoded into a 1-bit value for each state. Where the original vector is substituted, the new vector index is additionally encoded. For example, when the size of the codebook is less than 2,048, a vector index can be encoded into an 11-bit value. Therefore, where 10% of vectors are substituted, complementary information for F0 is packed to 10.5 (≈ 5 × 1 + 0.1 × 5 × 11) bits per phone.

It should be noted that only speech synthesis symbols and complementary information for duration and F0 are encoded into complementary information in this study. Spectral features are determined only from speech synthesis symbols and not modified by complementary information. Different from [6], state durations are not vector-quantized. Coding for spectral features and vector quantization of state durations will be examined in future work.

6. Evaluation

As the first study of the proposed method, the ability to reproduce the F0 contour of natural prosody was examined. To evaluate F0 errors, state durations of HMM are determined by forced alignment against target speech sounds. Since all of the durations are directly encoded into complementary information, the original durations are always reproduced.

6.1. Setup for evaluations

HMMs for speech synthesis were trained from 450 sentences in ATR503 phonetically-balanced sentences spoken by a female narrator. The sampling rate and precision were 16 kHz and 16 bits, respectively. HMMs were trained from 39-order mel-cepstral coefficients and log F0 that were extracted from the sounds every 5 milliseconds using HTS version 2.1[13]. The total numbers of leaf nodes of trained decision trees for mel-cepstrum and log F0 were 581 and 1,974, respectively.

6.2. Objective evaluation

For evaluation of the reproduction of the F0 contour, the remaining 53 sentences of ATR503 spoken by the same narrator were used. For the accuracy of speech synthesis symbols, symbols for the test set were transcribed by expert annotators.

Figure 3 illustrates the F0 contour of a sentence by the proposed method. The dotted, hatched, solid lines indicate target F0 extracted from the original speech sound, generated F0 by the conventional HMM-based speech synthesis and generated F0 by the proposed method, respectively. The horizontal lines at 130Hz in the figure indicate the periods of the states where the state feature vectors were substituted. In this figure, the state durations extracted from the original sound are used in both the conventional and proposed methods, and approximately 10% of vectors (22 vectors out of 225) for log F0 are substituted in the proposed method. The figure implies that the proposed method can appropriately reproduce the target F0 contour.

It should be noted that the contour is also significantly shifted for some frames not corresponding to substituted states. This is caused by extremely small variances of distributions for some states. Substitutions to distributions with small variances considerably affect other frames that do not directly correspond to substitution. Consequently, reproduction of natural prosody can be achieved by substitution of a small number of distributions.

Figure 4 shows the performance of the proposed method in the reproduction of the F0 contours. In the figure, the horizontal and vertical axis correspond to the substitution rate of the state feature vectors for F0 and RMSE of log F0 in semitone (= 1/12 octave), respectively. The results show that RMSE in log F0 can be reduced less than 0.3 semitones when 10% of F0 vectors are substituted, and RMSE can be reduced less than 0.3 semitone when 20% of vectors are substituted, an average of 53 sentences. Not only for the mean of 53 sentences, but also for the best 5 sentences and for the worst 5 sentences, the reduction of errors in F0 is saturated where the modification rate is approximately 20%. In this case, RMSE becomes approximately 15% of RMSE without the substitution of state feature vectors.

6.3. Subjective evaluation

To evaluate the proposed method, an ABX listening test was also conducted. In the test, 10 sentences out of the 53 sentences were evaluated. Stimuli A and B in the ABX test were the synthetic speech sound excited by F0 contours of the target speech sound, and the sound synthesized by the conventional HMM-based speech synthesis method without complementary information. Stimulus X was the sound by the proposed method with substitution of feature vectors for 5%, 10% or 20% of states. In the test, 8 subjects were asked whether A or B was similar to X. The total of 60 stimuli in the order of both A, B, X and B, A, X for the 10 sentences and the 3 conditions of the substitution were randomly presented for each subject.
In the result, the rates of selection of the speech sounds excited by the target F0 contours were 87.9%, 87.9%, and 88.6% in 5%, 10% and 20% substitution, respectively. The result confirms that the proposed method with small complementary information can reproduce natural F0 contours.

The stimuli by the proposed method that were not selected in the ABX test were preliminary investigated. For most of them, temporal structures in the spectra were distorted by inappropriate alignment of the states of the HMMs. In particular, such distortions caused by the states aligned with extremely long or short duration were frequently observed near prolonged vowels. Consequently, phonemically unnatural sounds were sometimes generated. In most of the stimuli, significant errors in F0 contours were also observed because modification by substitution of state feature vectors can be difficult especially in state with long duration. More appropriate alignment of the states will be eventually investigated in future studies for the vector substitutions of duration where extremely long or short state duration not included in the VQ codebook cannot appear. Use of the state-alignment directly taken into account for the number of repeated frames in each state, such as Hidden semi-Markov model (HSMM)-based forced-alignment supported in HTS-2.1.1, may also be improved such inappropriate alignment.

7. Conclusion

In this study, an extension method to reproduce natural prosody using HMM-based speech synthesizers was proposed. In the proposed method, only inappropriate state feature vectors of HMM for an utterance are substituted by other vectors in the decision trees for speech synthesis. Vectors for substitution are determined by a greedy search algorithm with a criterion based on the minimization in the error between the generated F0 contour and that of the target speech sound.

To evaluate the proposed method, an experiment with HMMs trained from 450 sentences spoken by a female narrator and 53 target sentences spoken by the same narrator was conducted. The results of the experiment indicated that 10% of the substitution of state feature vectors determined by speech synthesis symbols reduces error in log F0 to 0.4 semitones and 20% of substitution reduces error to 0.3 semitones. The result of the conducted subjective evaluation by the ABX test confirms that the proposed method can reproduce natural F0 contours.

In future work, not only reproduction of prosodic features, but also reproduction of spectral features will be examined. In addition, VQ of duration will also be examined for smaller complementary information.

8. References


