F_0 Modeling for Isarn Speech Synthesis using Deep Neural Networks and Syllable-level Feature Representation

Pongsathon Janyoi and Pusadee Seresangtakul
Department of Computer Science, Khon Kaen University, Thailand

Abstract: *The generation of the fundamental frequency (F_0) plays an important role in speech synthesis, which directly influences the naturalness of synthetic speech. In conventional parametric speech synthesis, F_0 is predicted frame-by-frame. This method is insufficient to represent F_0 contours in larger units, especially tone contours of syllables in tonal languages that deviate as a result of long-term context dependency. This work proposes a syllable-level F_0 model that represents F_0 contours within syllables, using syllable-level F_0 parameters that comprise the sampling F_0 points and dynamic features. A Deep Neural Network (DNN) was used to represent the relationships between syllable-level contextual features and syllable-level F_0 parameters. The proposed model was examined using an Isarn speech synthesis system with both large and small training sets. For all training sets, the results of objective and subjective tests indicate that the proposed approach outperforms the baseline systems based on hidden Markov models and DNNS that predict F_0 values at the frame level.*

Keywords: *Fundamental frequency, speech synthesis, deep neural networks.*

Received July 14, 2019; accepted May 28, 2020
https://doi.org/10.34028/iajit/17/6/9

1. Introduction

Statistical Parametric Speech Synthesis (SPSS) using Hidden Markov Models (HMMs) [30] uses statistical models to generate acoustic features instead of concatenating speech waveforms, thus generating human-like speech that is smoother and more stable than that produced by concatenative speech synthesis [39]. Its many advantages include the small data set required for training [28] and fixable voice characteristics [11]. Therefore, HMM-based speech synthesis systems are currently being utilized for several languages [1, 2, 17, 28].

The generation of fundamental Frequency (F_0) contours is an important issue in speech synthesis, as they have an influence on the naturalness of syntactic speech and human perception is very sensitive to unusual F_0 contours [5, 6]. F_0 contours are used to represent local tone and intonation. Intonation, which is represented by the global F_0 contour, is a major feature of accented languages such as English and Japanese. In these languages, tone might convey emotional information about the speaker, but indicates nothing about the meaning of the word. On the other hand, in tonal languages such as Thai and Mandarin Chinese, tone is a very important feature of syllables that is used to distinguish the meaning of words. Words with the same phoneme sequence, pronounced with different tone, have different meanings. Thus, tone correctness is crucial for speech synthesis systems for tonal languages. F_0 is normally used to identify the tone type. Typically, the F_0 contour of a monosyllabic word is stable, while the F_0 contour of a syllable in continuous speech is varied, owing to factors such as tone coarticulation, adjacent tones, and phrase intonation [6, 12, 21]

Several F_0 modeling approaches have been proposed. They can be divided into parametric model-based and frame-based approaches [13]. Parametric model-based approaches model the F_0 contour at the phone/syllable level. First, the F_0 contour within a segment is transformed into a set of parameters, then linguistic factors are mapped into the extracted parameters. Parameterization methods proposed to represent F_0 contours within syllables include the Tilt model [25, 27], pitch target approximation model [37], and target F_0 point [19]. F_0 contours can also be represented by the parameters of the Fujizaki model [5], which decomposes the F_0 contour into phrase components, and other representations such as coefficients from discrete-cosine transformation[18, 26]. Machine learning approaches, including decision-trees, support vector machines, and artificial neural networks, are used to learn the relationship between the linguistic factors and the extracted parameters. These models can represent the F_0 contour using a compact parameter set, and also have the advantage of the syllable being a more plausible unit of speech production than frames or phones [36]. Nevertheless, they require a fitting process to extract the set of parameters (e.g., the Fujizaki and pitch target approximation models), which can increase the complexity of the relationship between linguistic features and model parameters, for example, in the directed prediction of the Fujizaki model parameter in [10].

For frame-based approaches, the input linguistic features are directly mapped onto the F_0 value of each speech frame. For example, conventional HMM-based SPSS directly models the discontinuous F_0 using a multi-space probability distribution HMM Multi-Space Probability Distribution Hidden Markov Model (MSD-HMM) [29] to describe the distribution of the F_0 values in the voiced frames and the probability of being unvoiced. However, this approach considers F_0 in short intervals that are insufficient to deal with suprasegmental features of F_0 contours. Therefore, multi-layer F_0 models for HMM-based SPSS have been proposed [33, 34], using different F_0 models to model the pitch patterns for different prosodic layers. A simple tone-separated tree structure with contextual tone information was proposed to improve tone correctness of Thai HMM-based SPSS [3]. These approaches improve the naturalness of synthetic speech. However, the naturalness of synthetic speech is limited by the shortcoming of decision tree-based clustering that is insufficient to express complex context dependencies between linguistic features and acoustic features [41, 42].

In recent years, DNN-based SPSS was proposed [41], replacing decision trees with DNNs to model both F_0 and spectral features at the frame level. Several reports indicated that the DNN-based SPSS outperformed the HMM-based SPSS [9, 16]. DNNs are also employed to model F_0 values only, such as in the deep belief network-based F_0 model [15] and the continuous F_0 model [32].

However, conventional DNN-based systems cannot perfectly handle the complex variations of F_0 contours, because they consider them over a short interval such as a state or frame. To improve the F_0 modeling, a hierarchical F_0 model [38] has been proposed in which the F_0 contour is decomposed into sub-components, with each sub-component then individually modeled using contextual features at each prosodic layer. However, this model required more computational time than nonhierarchical versions. It also focused on modeling intonation rather than tone co-articulation and was applicable to accent languages because the F_0 contour of each segment was modeled without including contextual features at each level.

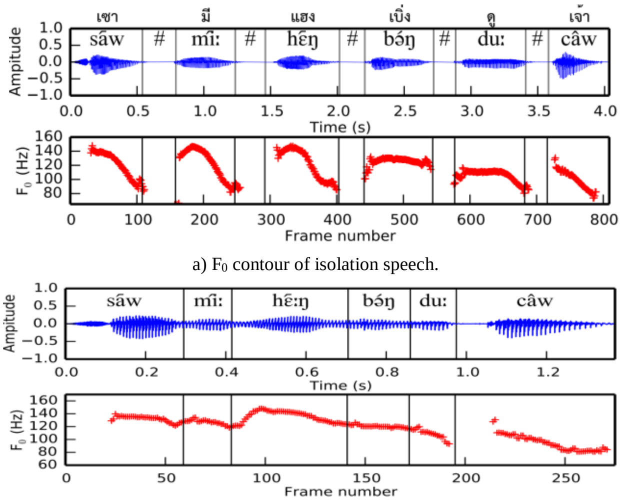
In speech synthesis for tonal languages, the deviation of F_0 contour is more associated with larger prosodic levels than frame and phone (i.e., syllables and words) due to tone co-articulation and intonation effects [6, 12, 21]. Therefore, the F_0 contour should be modeled at the syllable level than frame or phone.

In this work, we focused on improving tone modeling on DNN based speech synthesis for tonal languages. The proposed model was examined on SPSS for Isarn, a Thai dialect. We hypothesized that the representation of the F_0 contour within a syllable would provide more F_0 curve information than the frame-by-frame approach used in conventional SPSS. The F_0 contour was separately modeled by mapping between the contextual features and F_0 parameters of each syllable using a single DNN. Syllable-level F_0 parameters were estimated from all F_0 points of the syllable. Syllable-level dynamic features were added to preserve the sequence features of the current syllable and the adjacent syllables, and to generate a smooth F_0 contour. We also investigated the performance of the proposed method with limited training data by varying the size of the speech data set (e.g., 0.4, 1.2, 2.4, or 3.5 hours). The proposed model can be plugged into either HMM-based or DNN-based SPSS by including the generated F_0 contour. The use of DNN in this work does not greatly increase the computational cost, because the model generates the F_0 contour at the syllable level.

The rest of the paper is organized as follows. Section 2 briefly introduces the Isarn language, section 3 describes the proposed F_0 model, and section 4 describes the experiments and results. Our conclusions and recommendations for future study are presented in the final section.

2. Tone in Isarn Language

The Isarn language is a dialect of Thailand, which is widely used in the north-east region of the country [23]. Isarn can be written using the standard Thai alphabet. There are six tones in the Isarn language: mid, low, mid falling, high falling, high, and rising. Figure 1 shows a comparison of F_0 contours corresponding to Isarn utterances by a male native speaker, when pronounced in isolation (a) and spoken naturally in continuous speech (b). As shown in the top panel of Figure 1, the shapes of the F_0 contours of syllables in falling tone are quite similar.



a) F_0 contour of isolation speech.

b) F_0 contour of continuous speech.

Figure 1. Comparison of F_0 contours of Isarn utterance (“You should have some rest” in English translation) pronounced by the same speaker in isolation (a) and continuous speech (b).

The F_0 contour of a syllable uttered in isolation has a stable pattern. The stable pattern of each tone can be analyzed and generated using a parametric model [20]. In continuous speech, the F_0 patterns of syllables of the same tone slightly deviate from the stable pattern owing to complex linguistic factors that vary according to context. Previous studies reported that the main factors influencing the shape of the F_0 contour are tone co-articulation and intonation effects [6, 12, 21] . Tone co-articulation is a phenomenon in which the shape of F_0 contour of the current syllable is affected by the F_0 contours of the adjacent syllables, because the articulatory organs cannot respond rapidly enough to preserve the shape of the F_0 contours of uttered syllables. Intonation effects cause the F_0 contour of an utterance to gradually decline. The lower panel of Figure 1 shows how the F_0 contours of some syllables differ from the stable pattern, such as the occurrences of a high falling tone sequence. The shape of the F_0 contour of the current syllable is assimilated with the F_0 contour of the next syllable.

3. Proposed Model

Here, we propose a syllable-level model for generating F_0 contours for tonal language speech synthesis, to overcome the inability of frame-based models to represent the deviation of F_0 contours over long time periods [33, 40]. In tonal languages, tone is indicated by the F_0 contour at the syllable-level. The deviation of F_0 contours also occurs at the syllable level [21, 35]. Hence, we represent syllable-level F_0 contours by sampling F_0 values with dynamic features. A DNN is used to generate syllable-level F_0 parameters from syllable-level linguistic features, including features of the current and adjacent syllables to handle the deviation of F_0 contours due to tone coarticulation. Figure 2 shows an overview of the proposed model, comprising the F_0 modeling and generation parts.

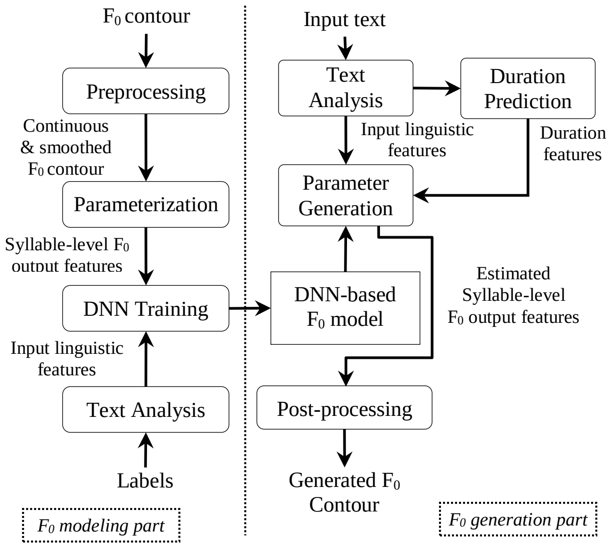


Figure 2. An overview of the proposed F_0 model.

3.1. F_0 Modeling

The F_0 modeling stage estimates the weights of the DNN used to represent the relationship between the linguistic features and F_0 contours, and consists of pre-processing, parameterization, and DNN training stages.

3.1.1. Preprocessing

Before modelling a given F_0 contour, interpolation and smoothing must be performed. The interpolation process fills artificial F_0 values between the intermediate F_0 values where there are unvoiced speech segments or short pauses. We first eliminated unusual F_0 values (e.g., error points and micro prosody) from regions of unvoiced speech based on the phoneme region described in the label files. The unvoiced speech segment was interpolated by the piecewise cubic interpolation method [4] and smoothed by the median filter method.

3.1.2. Parameterization

The length of a syllable depends on its phoneme identities and position in the utterance. Thus, we were unable to model the F_0 contour of a syllable using DNN directly, as DNN requires fixed dimensions of input and output features. Therefore, the smoothed F_0 contour of each syllable was scaled to a fixed dimension vector according to the algorithm 1.

Algorithm 1: Sampling $F_0(F_0,P,T,N,K)$

```
#Inputs:
 $F_0 = [f_1, f_2, f_3, \dots, f_T]$ is the extracted log $F_0$ vector
 $T$ is the number of frame in the utterance.
$N$ is the number of syllable in the utterance.
$K$ is the number of sampled $F_0$ values per syllable.
$P=[(b_1,e_1), (b_2,e_2), (b_3,e_3),\dots, (b_N,e_N)]$ is the list of start and end frame positions of syllable.
#Output
$C$ is the scaled log $F_0$ vector of length $K \times N$
for $i = 1$ to $N$ do
        $(b,e) \leftarrow P[i]$
        for $j = 1$ to $K$ do
                $C[(i-1)K+j] \leftarrow F_0[1+b+(j-1)(e-b)/K]$
        end for
end for
return $C$
```

However, using only static features can generate mismatched F_0 contours at the connection point of two syllables, because the F_0 vector of each syllable is generated individually. Thus, dynamic features were taken as part of the output features, to preserve the continuation of F_0 contours of the current and adjacent syllables, and to improve the smoothness of the F_0 contour at the connection points between syllables. The dynamic features were calculated based on a conventional algorithm [43] used in both conventional HMM-based and DNN-based SPSS. In this step, the scaled F_0 vector with dynamic features C_d contains a sequence of sampled F_0 values, including the delta and delta-delta, as follows:

$$C_d = [C \quad \Delta C \quad \Delta\Delta C] \tag{1}$$

$$= [CW_0 \quad CW_1 \quad CW_2] \tag{2}$$

Where C is the static feature vector of the scaled log F_0 sequence and W_n is a window matrix for calculating the n-th dynamic features of the scaled log F_0 sequence.

In the final step, the C_d is reshaped as the output feature in order to model F_0 values at the syllable level, as shown in the algorithm 2.

Algorithm 2 ReshapeOutputFeature(C_d,N,K)

```
#Input:
C_d is the matrix of scaled log F_0 values with dynamic features.
N is the number of syllable in the utterance.
K is the number of sampled F_0 values per syllable.
#Output
Y is an $N \times 3K$ zero matrix for storing the output features.
for i = 1 to N do
        $Y[i,1:K] \leftarrow C_d[iK:(i+1)K,1]^T$
        $Y[i,K:2K] \leftarrow C_d[iK:(i+1)K,2]^T$
$Y[i,2K:3K] \leftarrow C_d[iK:(i+1)K,3]^T$
end for
return Y
```

3.1.3. DNN Training

For an utterance of N syllables, $X=[x_1\ x_2\ \dots\ x_N]$ is the sequence of the input features generated using the linguistic specification, and $Y=[y_1\ y_2\ \dots\ y_N]$ is the sequence of output features from the extracted F_0 contour. The DNN maps the input feature vectors x_i to the output feature vectors y_i, where i is the syllable index [7] . For each unit at each hidden layer, the input from the layer below is mapped to a deterministic value using a nonlinear activation function σ(.) and passed to the layer above. This can be expressed as:

$$h^{(0)} = x_i \tag{3}$$

$$h^{(l)} = \sigma\left(W^{(l)} h^{(l-1)} + b^{(l)}\right) \tag{4}$$

$$y_i = h^{(L)} \tag{5}$$

Where x_i is the input feature vector of syllable i, y_i is the output feature vector of syllable i, $h^{(l)}$ is the output of hidden layer l, $h^{(0)}$ is the input layer, $W^{(l)}$ is a weight matrix for the link between the layer l-1 and l, L denotes the number of layers of the network, and $b^{(l)}$ is the bias vector of the hidden layer l.

To model the F_0 parameters at the syllable level, we modified the phone-level linguistic features used in our previous HMM-based speech synthesis system [8] by replacing the quinphone features with the articulatory features of the phonemes within the current syllable. The input linguistic features are listed in Table 1. For feature extraction, the features of the previous syllable and the next syllable were included in addition to those of the current syllable. These features can be divided into four categories: tone features, duration features, articulatory features, and positional features. This input feature set was designed in order to use tone features to describe the global pattern of tone sequence; other feature sets were used to represent the deviation of F_0 patterns.

The articulatory features of the syllable can be divided into phoneme identity and phoneme category features. The phoneme identity features represent the sequence of sound units in a syllable (e.g., /b/, /a/, /n/), whereas the phoneme category features identify the type of sound units in the syllable (e.g., nasal, plosive, fricative). We included phoneme category features because it was not possible to predict all possible phoneme identities of syllables, and we hypothesized that the phoneme category would be more useful than the phoneme identity for generating the F0 parameters of unknown syllables.

Table 1. Input linguistic features for training the proposed model.

Feature categories	**Description**	**#dims.**
Tone features	Tone identities of previous / current / next syllables.	23
Duration feature	Duration of previous / current / next syllables.	3
	Duration of initial consonant/ vowel / final consonant of current syllable.	3
Articulatory features	Phoneme identities of initial consonant/ vowel / final consonant of current syllable.	50
	Phoneme categories of initial consonant/ vowel / final consonant of current syllable.	32
	Phoneme identities of initial consonant and vowel of next syllable.	42
	Phoneme identities of vowel and final consonant of previous syllable.	29
	Phoneme categories of initial consonant and vowel of next syllable.	29
	Phoneme categories of vowel and final consonant of previous syllable.	16
Positional features	Position of the current syllable in the current word.	1
	Number of syllables in the current word.	1
	Position of the current syllable in the current phrase.	1
	Number of syllables in the current phrase.	1
	Position of the current word in the current phrase.	1
	Number of words in the current phrase.	1
	Position of the current phrase in the current phrase.	1
	Number of words in the current phrase.	1
	Syllable section in the current word (silence, single, begin, middle, end).	5
	Word section in the current phrase (silence, single, begin, middle, end).	5
	Phase section in the utterance (silence, single, begin, middle, end).	5

We represented the symbolic features using the one-hot representation method. For example, a seven-dimensional vector was used to represent the tone identities of the current syllable because there are six tones in Isarn in addition to the pause symbol. Numeric features such as the duration of the syllable and its position in the current phrase were directly used in the training process.

3.2. F_0 Generation Process

To generate the F_0 contour from input text, the input text is converted to linguistic features using the text analysis module, and the linguistic features are used to predict the duration of phonemes using DNN-based duration model. Next, the duration features are included in the linguistic features, and the input features are fed into the trained DNN, to obtain the predicted output features. Then, the non-smooth F_0 vector with its dynamic features is reshaped as a single vector using the inversion of the algorithm 2, and the F_0 contour is generated using a parameter generation algorithm [31]. Finally, the smoothed F_0 vector is scaled to the duration of the syllable, obtained using the duration predictor.

4. Experiments and Results

4.1. Speech Corpus and Feature Extraction

The experiments in this work employed an Isarn speech corpus containing 4,400 utterances of a male native speaker, to conduct Isarn HMM-based speech synthesis system [8]. Training sets of different sizes were used to investigate the performance of the proposed model with limited training data. We created four training sets, comprising T0.4: 40 minutes (500 utterances), T1.2: 1 hour 20 minutes (1,000 utterances), T2.4: 2 hours 40 minutes (2,000 utterances), and T3.5: 3 hours 50 minutes (3,000 utterances).

The development and evaluation sets were both 20 minutes (200 sentences) in duration. Statistical information for the T3.5 training set is given in Table 2. The total number of samples of each tone is shown in Table 3. We used a sampling rate of 32,000 Hz instead of the rate of 16,000 Hz used in some other works, as this does not degrade the quality of syntactic speech and is equivalent to using higher sampling rates [24].

These acoustic features were extracted using the WORLD vocoder [14] with a 5 ms frame shift. The output acoustic features consisted of four parts: the spectral envelope Mel Cepstral Coefficient (MCC), Band Aperiodicity (BAP), log F_0, and voiced/unvoiced (V/UV) flag.

Table 2. Statistical information of the speech data of a male native speaker.

Description	**Values**
Number of syllables	46,251
Number of words	37,761
Number of phrases	5,151
Length of utterance (phrase)	1.91 ± 0.78
Length of phrase (syllable)	11.10 ± 5.00
Length of word (syllable)	1.21 ± 0.48
F_0 (Hz)	113.25 ± 20.09

Table 3. Total number of samples of each tone.

Tone	**Number of samples**
Mid tone	10,411
Low tone	7,813
Mid falling tone	7,165
High falling tone	8,721
High tone	5,800
Rising tone	6,341

We needed to determine the appropriate number of sampled F_0 values per syllable, because this number could affect the accuracy of the model. With a small number of F_0 values (e.g., 3 or 5), the detail of the tone curve may deviate, especially in the case of long syllables. On the other hand, a large number of F_0 values (e.g., 100 or 200) would require fabrication of unnecessary and redundant F_0 values for short syllables. Thus, we set the number of sampled F_0 values per syllable to 40, approximately the average number of frames of all syllables in the speech database. This number is determined by considering the distribution of syllable duration in the speech corpus as shown in Figure 3.

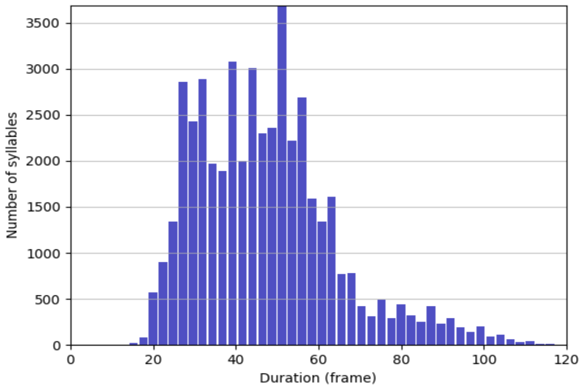


Figure 3. Distribution of syllable duration in the speech corpus.

4.2. Experimental Setup

4.2.1. Speech Generation

In SPSS, various spectral parameters in addition to F_0 are required for generating syntactic speech. The model presented in this work produces only F_0; thus, additional models for generating the other speech parameters (MCC, BAP, and V/UV) were incorporated. In this work, DNN is employed to generate these speech parameters. For DNN training, the input vectors consisted of 258-dimensional linguistic features, containing 250 binary values and eight numerical values. We used the coarse-code as frame-level feature, similar to the original work [41]. Each observation vector consisted of 60 MCCs, log F_0, four BAPs, their delta and delta-delta features, and a voiced/unvoiced binary flag. The input features were normalized to the range of [0.01, 0.99], and the output features were normalized by mean and variance. The number of hidden layers was varied from 2 to 8 and the number of hidden nodes was varied in set of [128, 256, 512, 1024, 2048]. The DNNs was trained using Adam optimization algorithm with α=0.0001, β_1=0.9, β_2=0.999, ε=10^{-8}, a batch size of 128, a learning rate of 10^{-4}. To avoid over-fitting, we applied early stopping criteria to stop training when the validation loss has stopped decreasing in 20 epochs.

4.2.2. Comparison of Systems

To evaluate the performance of the proposed F_0 model, the HMM-based and DNN-based F_0 models were taken as the baseline. The details of each system can be summarized as follows.

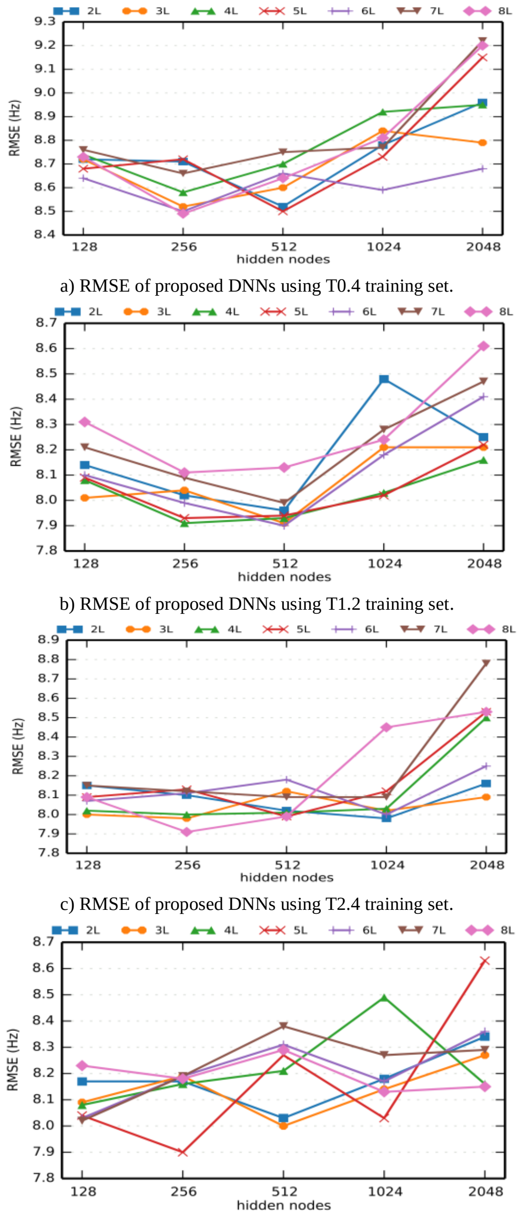
- Baseline HMM: For the training of HMM-based speech synthesis, a left-to-right, no-skip hidden semi-Markov model with single mixture and a diagonal covariance matrix was used to model the acoustic features. The spectral part was modeled by a Continuous Probability Distribution (CD), and F_0 was modeled by MSD-HMM. The state-level decision trees were constructed from the states of the context-dependent HMMs by using the Minimum Description Length (MDL) criterion [22] as the stopping criterion for splitting decision trees.
- Baseline DNN: DNNs was trained for predicting only F_0 frame by frame because it has more parameters to model the F_0 contour than the use of DNN to model all speech parameters simultaneously. The DNN was trained with the same input features that were used for training the DNN-based SPSS described in subsection 4.2.1. The output features of the DNN consisted of log F_0, the delta and delta-delta features, and a voiced/unvoiced binary flag. The hyper-parameters were tuned similar to training the DNN-based SPSS.

4.2.3. Optimization of DNN Parameters

To conduct a meaningful comparison between baseline systems and the proposed model, we needed to select the highest-performing model. In this work, the proposed model and baseline model were trained by varying the number of hidden nodes and layers. The number of hidden layers was varied from two to eight, and the number of hidden nodes was varied in the set of [128, 256, 512, 1024, 2048]. For the proposed model, the input vectors of the DNNs consisted of 250-dimensional linguistic features, containing 242 binary values and eight numerical values. The 40-dimensional F_0 vector and its dynamic features were taken as the output features. In total, the output features consisted of 120 numerical values. A tanh activation function was used in the hidden layers, followed by linear activation at the output layer. For the training procedure, the weights of the DNN were initialized randomly, then optimized to minimize the mean squared error between the output features of the training data and predicted values, using the Adam-based back-propagation algorithm. The parameters for the Adam algorithm were set as α=0.00001, β_1=0.9, β_2=0.999, and ε=10^{-8}. Figure 4 presents the RMSE of the proposed DNNs for each training set. The best-performing DNNs for the T0.4, T1.2, T2.4, and T3.5 training sets were $\text{DNN}_{8\times256}$, $\text{DNN}_{6\times512}$, $\text{DNN}_{8\times256}$, and $\text{DNN}_{5\times256}$, respectively. Based on these results, the best-performing DNN for each training set had more than four hidden layers, and 256 or 512 hidden nodes.

4.3. Objective Tests

The best-performing DNNs in terms of objective results for each system and each training set are listed in Table 4. The RMSE of the baseline DNN and proposed DNN were clearly lower than that of the baseline HMM, and the baseline DNN gave a lower RMSE than baseline HMM. These results were contrary to those of a previous study [41], possibly because the tone type included in the input feature can enhance the performance of the DNN. Comparing the baseline DNN and proposed DNN, we found that the RMSE of the proposed DNN was slightly lower than that of the baseline DNN.



a) RMSE of proposed DNNs using T0.4 training set.

b) RMSE of proposed DNNs using T1.2 training set.

c) RMSE of proposed DNNs using T2.4 training set.

d) RMSE of proposed DNNs using T3.5 training set.

Figure 4. RMSE of the proposed DNNs trained using the different numbers of hidden layers and nodes for T0.4 (a), T1.2 (b), T2.4 (c) and T3.5 (d) training sets (#L denotes the number of hidden layers).

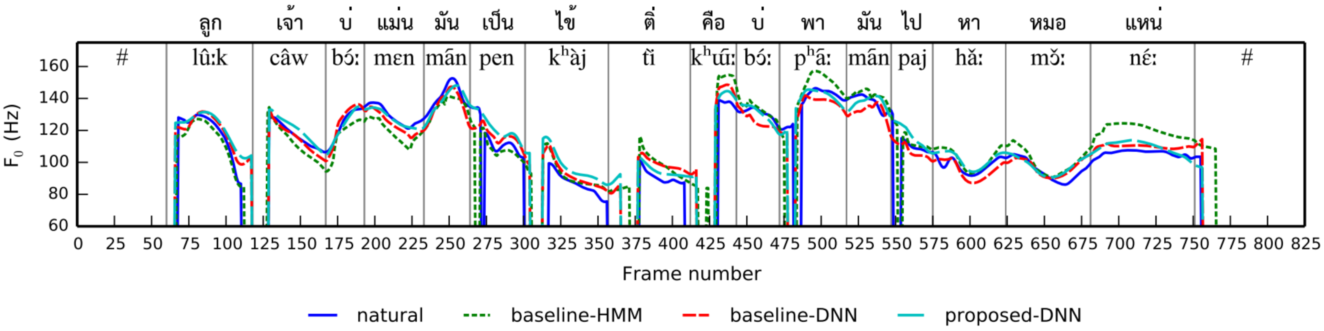


Figure 5. Comparison of F0 contours generated using baseline-HMM, baseline-DNN, and proposed DNN.

Table 4. Objective result of the baseline systems and the proposed F0 model ($DNN_{\#L\times\#H}$: #L and #H represent the number of hidden layers and nodes, ± denotes standard deviation).

Training set	System	RMSE (Hz)	CORR
T0.4	**baseline HMM**	10.03 ± 2.92	0.89 ± 0.06
	baseline $DNN_{7\times512}$	8.60 ± 2.70	0.91 ± 0.05
	proposed-$DNN_{8\times256}$	**8.49 ± 2.97**	**0.92 ± 0.05**
T1.2	**baseline HMM**	9.46 ± 2.69	0.89 ± 0.06
	baseline $DNN_{5\times512}$	8.03 ± 2.45	0.92 ± 0.04
	proposed-$DNN_{6\times512}$	**7.90 ± 2.47**	**0.93 ± 0.04**
T2.4	**baseline HMM**	9.54 ± 2.89	0.90 ± 0.06
	baseline $DNN_{7\times512}$	8.09 ± 2.64	0.93 ± 0.04
	proposed-$DNN_{8\times256}$	**7.91 ± 2.69**	**0.93 ± 0.05**
T3.5	**baseline HMM**	9.76 ± 3.09	0.90 ± 0.06
	baseline $DNN_{7\times256}$	8.22 ± 2.82	0.93 ± 0.04
	proposed-$DNN_{5\times256}$	**7.90 ± 2.72**	**0.93 ± 0.04**

4.4. Subjective Tests

We verified the objective results by conducting preference listening tests to compare the perceptual quality of synthetic speech generated using the baseline HMM, baseline DNN, and proposed DNN. As these tests were used to investigate the F_0 generation performance of the systems, the spectral parameters were generated using the same system, whereas the F_0 was generated using different models.

Twenty sentences were selected from the test set, and 20 native listeners participated in this test. In this test, the original phone duration was used in order to avoid the adverse effects of the duration model. To reduce noise caused by fixing the order of choices, 20 test sentence pairs were selected and the order of two samples in each pair was swapped. As a listener can become exhausted after listening for a long time, potentially affecting their perception performance, listeners were asked to play each sample once or twice.

Three preference tests were conducted for each training set, consisting of comparisons between the baseline HMM and proposed DNN, between the baseline DNN and proposed DNN, and between the baseline HMM and baseline DNN. Table 5 shows the preference scores for the proposed system compared with those of baseline HMM and DNN. A *t*-test was used to show that the differences between the compared systems were significant ($p < 0.01$). As shown in the table, the baseline DNN and proposed DNN scored higher than the baseline HMM, and the preference score of the proposed DNN was higher than that of the baseline DNN, indicating that the proposed system significantly outperformed both the baseline HMM and the baseline DNN. The results of the subjective tests were similar to those of the objective tests.

Table 5. Subjective preference score (%) of each system pairs. N/P denotes no preference.

Training set	System			N/P	*p*-value
	baseline HMM	baseline DNN	proposed DNN		
T0.4	30.56	**59.17**	-	10.28	0.001
	22.50	-	**71.94**	5.56	$< 10^{-10}$
	-	35.00	**55.58**	9.47	$< 10^{-8}$
T1.2	29.17	**63.06**	-	7.78	$< 10^{-6}$
	21.17	-	**71.87**	6.96	$< 10^{-8}$
	-	32.87	**63.51**	3.62	$< 10^{-5}$
T2.4	26.48	**66.48**	-	7.04	$< 10^{-5}$
	21.39		**70.28**	8.33	$< 10^{-8}$
	-	39.17	**52.78**	8.06	0.004
T3.5	27.22	**67.78**	-	5.00	$< 10^{-8}$
	23.61	-	**71.94**	4.44	$< 10^{-9}$
	-	33.43	**60.72**	5.85	$< 10^{-7}$

A comparison of the F_0 contours generated by three systems is shown in Figure 5. The F_0 contours are generated from the Isarn sentence ("Your child may be sick, he should go to see the doctor" in English translation). As shown in the figure, the proposed system could generate the precise shape of the F_0 contour within a syllable better than baseline systems (e.g., from frame 375 to frame 410, and from 450 to 550).

5. Conclusions and Future Work

Here, we presented an F_0 model for tonal language SPSS, which generates F_0 contours at the syllable-level using a DNN. The proposed model was examined on an Isarn SPSS, using both small and large training sets, and evaluated using objective and subjective tests, to determine the precision of F_0 prediction and perceptual quality, respectively. Baseline HMM and DNN systems were compared with the proposed model. Both objective and subjective test results suggested that the proposed model outperformed baseline HMM and DNN systems for all training sets. However, this work focused only on F_0 modeling, whereas the prosody of speech involves both the F_0 contour and duration. In future work, a duration model will be incorporated into the current system.

Acknowledgments

This work was supported through a Computer and Information Science Interdisciplinary Research Grant from the Department of Computer Science, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand.

References

[1] Amrouche A., Falek L., and Teffahi H., "Design and Implementation of a Diacritic Arabic Text-To-Speech System," *The International Arab Journal of Information Technology*, vol. 14, no. 4, pp. 488-494, 2017.

[2] Chomphan S. and Kobayashi T., "Implementation and Evaluation of an HMM-Based Thai Speech Synthesis System," *in Proceedings of 8th Annual Conference of the International Speech*, Antwerp pp. 2849-2852, 2007.

[3] Chomphan S. and Kobayashi T., "Tone Correctness Improvement in Speaker Dependent HMM-Based Thai Speech Synthesis," *Speech Communication*, vol. 50, no. 5, pp. 392-404, 2008.

[4] Fujisaki H., Narusawa S., and Maruno M., "Pre-Processing of Fundamental Frequency Contours of Speech for Automatic Parameter Extraction," *in Proceedings of International Conference on Signal Processing*, Beijing, pp. 722-725, 2000.

[5] Fujisaki H. and Hirose K., "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," *Journal of the Acoustical Society of Japan*, vol. 5, no. 4, pp. 233-242, 1984.

[6] Gandour J., Potisuk S., and Dechongkit S., "Tonal Coarticulation in Thai," *Journal of Phonetic*s, vol. 22, no. 4, pp. 477-492, 1994.

[7] Goodfellow I., Bengio Y., and Courville A., *Deep Learning*, MIT Press, 2016.

[8] Janyoi P. and Seresangtakul P., "An Isarn Dialect HMM-Based Text-To-Speech System," *in Proceedings of 2nd International Conference on Information Technology*, Nakhonpathom, pp.1-6, 2017.

[9] Lazaridis A., Potard B., and Garner P., "DNN-Based Speech Synthesis: Importance of Input Features and Training Data," *in Proceedings of International Conference on Speech and Computer*, Athens, pp. 193-200, 2015.

[10] Li Y., Tao J., Hirose K., Xu X., and Lai W., "Hierarchical Stress Modeling and Generation in Mandarin for Expressive Text-To-Speech," *Speech Communication*, vol. 72, pp. 59-73, 2015.

[11] Masuko T., Tokuda K., Kobayashi T., and Imai S., "HMM-Based Speech Synthesis with Various Voice Characteristics," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2760, 1996.

[12] Mittrapiyanuruk P., Hansakunbuntheung C., Tesprasit V., and Sornlertlamvanich V., "Issues in Thai Text-to-Speech Synthesis: the NECTEC Approach," *NECTEC Technical Journal*, vol. 2, no. 7, pp. 36-47, 2000.

[13] Mnasri Z., Boukadida F., and Ellouze N., "F_0 Contour Modeling for Arabic Text-to-Speech Synthesis Using Fujisaki Parameters and Neural Networks," *Signal processing: An International Journal*, vol. 6, no. 4, pp. 352-369.

[14] Morise M., Yokomori F., and Ozawa K., "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877-1884, 2016.

[15] Mukherjee S. and Mandal S., "F_0 Modeling In Hmm-Based Speech Synthesis System Using Deep Belief Network," *in Proceedings of 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques*, Phuket, pp. 1-5, 2014.

[16] Qian Y., Fan Y., Hu W., and Soong F., "On the Training Aspects of Deep Neural Network (DNN) for Parametric TTS Synthesis," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, pp. 3829-3833, 2014.

[17] Qian Y., Soong F., Chen Y., and Chu M., "An HMM-Based Mandarin Chinese Text-To-Speech System," *in Proceedings of International Symposium on Chinese Spoken Language Processing*, Singapore, pp. 223-232, 2006.

[18] Ribeiro M. and Clark R., "A Multi-Level Representation of F_0 Using the Continuous Wavelet Transform and the Discrete Cosine Transform," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, pp. 4909-4913, 2015.

[19] Sagisaka Y., "On the Prediction of Global F_0 Shape for Japanese Text-To-Speech," *in Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 325-328, 1990.

[20] Seresangtakul P. and Takara T., "Analysis of Pitch Contour of Thai Tone Using Fujisaki's Model," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, pp. 505-508, 2002.

[21] Seresangtakul P. and Takara T., "Synthesis of Polysyllabic Sequences of Thai Tones Using a Generative Model of Fundamental Frequency Contours," *IEEJ Transactions on Electronics Information and Systems*, vol. 125, no. 7, pp. 1101-1108, 2005.

[22] Shinoda K. and Watanabe T., "MDL-Based Context-Dependent Subword Modeling for Speech Recognition," *Acoustical Science and*

Technology, vol. 21, no. 2, pp. 79-86, 2000.

[23] Siriaksornsat P., Thai Dialects. Bangkok: Department of Thai and Oriental Languages, Ramkhamhaeng University, 2011.

[24] Stan A., Yamagishi J., King S., and Aylett M., "The Romanian Speech Synthesis (RSS) Corpus: Building A High Quality HMM-Based Speech Synthesis System Using A High Sampling Rate," *Speech Communication*, vol. 53, no. 3, pp. 442-450, 2011.

[25] Taylor P., "Analysis and synthesis of Intonation Using The Tilt Model," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697-1714, 2000.

[26] Teutenberg J., Watson C., and Riddle P., "Modelling and Synthesising F_0 Contours with the Discrete Cosine Transform," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, pp. 3973-3976, 2008.

[27] Thangthai A., Thatphithakkul N., Wutiwiwatchai C., Rugchatjaroen A., and Saychum S., "T-Tilt: A Modified Tilt Model for F_0 Analysis and Synthesis in Tonal Languages," *in Proceedings of 9th the Annual Conference of the International Speech Communication Association*, Brisbane, pp. 2270-2273, 2008.

[28] Tokuda K., Black A., and Zen H., "An HMM-Based Speech Synthesis System Applied to English," *in Proceedings of IEEE Workshop on Speech Synthesis*, Santa Monica, pp. 227-230, 2002.

[29] Tokuda K., Masuko T., Miyazaki N., and Kobayashi T., "Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, pp. 229-232, 1999.

[30] Tokuda K., Nankaku Y., Toda T., Zen H., Yamagishi J., and Oura K., "Speech Synthesis Based on Hidden Markov Models," *in Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234-1252, 2013.

[31] Tokuda K., Yoshimura T., Masuko T., Kobayashi T., and Kitamura T., "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, pp. 1315-1318, 2000.

[32] Tóth B. and Csapó T., "Continuous Fundamental Frequency Prediction With Deep Neural Networks," *in Proceedings of 24th European Signal Processing Conference*, Budapest, pp. 1348-1352, 2016.

[33] Wang C., Ling Z., Zhang B., and Dai L., "Multi-Layer F_0 Modeling for HMM-Based Speech Synthesis," *in Proceedings of 6th International Symposium on Chinese Spoken Language Processing*, Kunming, pp. 1-4, 2008.

[34] Wu Y. and Soong F., "Modeling Pitch Trajectory by Hierarchical HMM with Minimum Generation Error Training," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, pp. 4017-4020, 2012.

[35] Wutiwiwatchai C., Hansakunbuntheung C., Rugchatjaroen A., Saychum S., Kasuriya S., and Chootrakool P., "Thai Text-to-Speech Synthesis: A Review," *Journal of Intelligent Informatics and Smart Technolo*gy, vol. 2, no. 2, pp. 1-8, 2017.

[36] Xu Y., "Speech Melody As Articulatorily Implemented Communicative Functions," *Speech Communication*, vol. 46, no. 3, pp. 220-251, 2005.

[37] Xu Y. and Wang Q., "Pitch Targets and Their Realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, no. 4, pp. 319-337, 2001.

[38] Yin X., Lei M., Qian Y., Soong F., He L., Ling Z., and Dai L., "Modeling F_0 Trajectories in Hierarchically Structured Deep Neural Networks," *Speech Communication*, vol. 76, no. C, pp. 82-92, 2016.

[39] Yoshimura T., Tokuda K., Masuko T., Kobayashi T., and Kitamura T., "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis," *in Proceedings of 6th European Conference on Speech Communication and Technology*, Budapest, pp. 2347-2350, 1999.

[40] Yu K., "Review of F_0 Modelling and Generation in HMM Based Speech Synthesis," *in Proceedings of IEEE 11th International Conference on Signal Processing*, Beijing, pp. 599-604, 2012.

[41] Ze H., Senior A., and Schuster M., "Statistical Parametric Speech Synthesis Using Deep Neural Networks," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, pp. 7962-7966, 2013.

[42] Zen H., Tokuda K., and Black A., "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.

[43] Zen H., Tokuda K., and Kitamura T., "Reformulating the HMM as A Trajectory Model by Imposing Explicit Relationships Between Static and Dynamic Feature Vector Sequences," *Computer Speech and Language*, vol. 21, no. 1, pp. 153-173, 2007.

Pongsathon Janyoi received the B.Sc. degree and M.S. degree in Computer Science from Khon Kaen University, Khon Kaen, Thailand, in 2010 and 2015, respectively. Currently, he is a Ph.D. candidate in Natural Language and Speech Processing Laboratory, Department of Computer Science, Khon Kaen University. His current research interests include speech synthesis, automatic speech recognition and machine learning.

Pusadee Seresangtakul received B.Sc. in Physics (Khon Kaen University) and M.Sc. in Computer Science (Chulalongkorn University), Thailand in 1986, and 1991, respectively. In 2005, she received a Ph.D. in Interdisciplinary Intelligent Systems Engineering from Graduate School of Engineering and Science, the University of the Ryukyus, Japan. She is currently an assistant professor in the Department of Computer Science, Faculty of Science, Khon Kaen University. Her research interests include NLP, speech processing, machine learning, and artificial intelligence system.