TRANSFORMATION OF F0 CONTOURS FOR LEXICAL TONES IN CONCATENATIVE SPEECH SYNTHESIS OF TONAL LANGUAGES

Trung-Nghia Phung [1], *Mai Chi Luong* [2], *Masato Akagi* [1]

[1] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[2] Institute of Information Technology, Hanoi, Vietnam

ABSTRACT

Concatenative speech synthesis (CSS) provides the greatest naturalness. However, it requires a huge stored database resulting a huge footprint. Reducing the capacity of stored database while preserving the quality of CSS, or improving the *quality to size ratio (QSr)*, is still a challenge. In this paper, we propose a method of transforming fundamental frequency (F0) contours of lexical tones, developed from TD-GMM framework that successfully applied for transforming spectral sequence in previous researches, in order to improve the *QSr* of CSS of tonal languages that results CSS available with limited data at offline stage, storing small online footprint, while preserving perceptual quality. The experimental results show that the proposed F0 transformation outperforms conventional and state-of-the-art F0 contour transformations for transforming lexical tones in terms of speech quality. When applying the proposed F0 contour transformation for transforming lexical tones in CSS of tonal languages, the *QSr* is enhanced compared with the method of simple F0 exchange while the quality of synthetic speech is preserved.

Index Terms— Concatenative, speech synthesis, tone transformation, quality to size ratio.

1. INTRODUCTION

Concatenative speech synthesis is the most natural speech synthesis up to now. However, the stored offline database is required to be huge resulting a huge online footprint of the synthesis [1]. This drawback limits practical applications of CSS. Reducing the size of stored database while preserving quality of CSS is therefore an interesting issue.

In the literature, Kain et al. [2, 3] attempts to improve the *QSr* instead of improving only quality, in which the larger the *QSr*, the more successful the CSS. To improve the *QSr*, they attempt to eliminate the mismatched-context errors in order to reduce the amount of concatenation data in CSS with a small reduction on speech quality. The approach proposed by Kain et al. is general, however for specific languages, it may have some other approaches motivating the same target of improving the *QSr* of CSS.

This research motivates to improve the *QSr* of CSS of tonal languages, resulting CSS available with limited data at offline stage, storing small online footprint, and still preserve perceptual quality. To achieve this aim, we propose an F0 transformation to convert lexical tones.

F0 transformations have been studied and applied in many researches [4, 5] but the number of studies on F0 transformation for lexical tones is surprisingly small [6]. Therefore, this research firstly investigates conventional and state-of-the-art F0 transformations in terms of converting lexical tones. Secondly, we investigate the temporal decomposition - Gaussian mixture model (TD-GMM) framework that successfully used for transformation of spectral sequence [8] to modify and optimize it with the proposed F0 transformation for converting lexical tones. Considering these F0 transformations and TD-GMM framework, this paper proposes two methods in order to propose an efficient F0 transformation for lexical tones. One is a modification of TD-GMM framework for converting dynamic F0 features close with lexical tones. Another one is a non-parallel alignment and training method that requires small offline data, used instead of phone-based alignment and training in original TD-GMM framework that requires a huge offline data to train all phone-based GMMs.

We evaluate the proposed method compared with simple F0 exchange [6], conventional F0 transformation [4] and state-of-the-art F0 transformation [5] for transforming lexical tones in terms of speech quality. After that, we evaluate the proposed method compared with simple F0 exchange in terms of the *QSr*, to confirm whether the proposed F0 transformation improves the quality and the *QSr* of CSS of tonal languages.

2. F0 TRANSFORMATIONS FOR LEXICAL TONES

2.1. Using Tone Transformations in CSS of Tonal Languages

It is known that the number of phonetic units in tonal languages is usually much larger than that in verbal languages, resulting the required concatenation data for conventional CSS of tonal languages is usually much larger than that in verbal languages. Therefore, reducing the number of tonal units needed to be stored for concatenation is an important issue.

Changing the tone for each pronunciation provides a set of tonal units that we call same-phonation set. F0

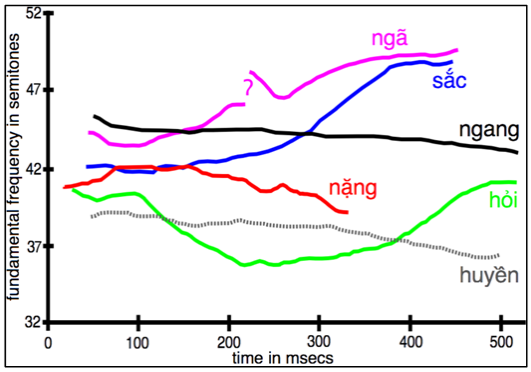

Fig. 1. General F0 contours of Six Vietnamese Tones: tone 1 (ngang - level), tone 2 (huyền - falling), tone 3 (ngã - broken), tone 4 (hỏi - curve), tone 5 (sắc - rising) and tone 6 (nặng - drop), adopted from [9].

transformation can be applied in CSS of tonal languages by combining the transformed F0 contours of tonal units with the spectral envelope of a representative unit in a same-phonation set to produce synthetic tonal units with a source/filter coder. The popular representative unit in a same-phonation set is the neutral unit with neutral tone, which is the tone with flat F0 contour usually existed in tonal languages [7]. Assume that all tonal syllables are converted instead of storing the original ones and denote the theoretical data reduction percentage as r_f. Then, r_f can be approximately computed as in Eq. (1),

$$r_f = (1 - N_n/N_t) \times 100\% \tag{1}$$

where N_n is the number of neutral syllables, N_t is the number of tonal syllables.

There are totally about 6776 meaningful tonal syllables and 1287 neutral syllables in Vietnamese [12]. Thus, $r_t \approx 81\%$ with Vietnamese. As a result, transforming F0 contour of lexical tones is an interesting method to reduce the required data for CSS. However, to improve the *QSr*, we need not only reduce the amount of required data but also improve the quality. The more high-quality transformed speech, the more accuracy of the F0 transformation of lexical tones is required. Therefore, in the next sub-sections, we investigate problems of conventional and state-of-the-art F0 transformations to accurately transform the lexical tones.

2.2. Simple Exchange of F0 Contours

The simplest F0 transformation for converting lexical tones is simple exchange the stored F0 contour of a tonal unit with F0 contour of a representative unit. This method has been efficiently applied for Thai language and reduces about five times of footprint [6]. However, this method requires a huge data covering all tonal units to compute F0 contour of all tonal units at offline stage. Thus, this method is still not available when having only small corpora and the *QSr* of this methods can be improved more.

2.3. Transformation of F0 Mean and Variance

The conventional F0 transformation approach is transforming the mean and variance of the F0 (logF0) distribution of the source speech to match those of the target speech [4]. Denoting the F0 value of a single frame of the source speech by x, the converted F0 value $\hat{y}$ is obtained as given in Eq. (2).

$$\hat{y} = \frac{\sigma_y}{\sigma_x}(x - \mu_x) + \mu_y \tag{2}$$

where μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of the training data for the source and the target speech, respectively. This method only changes the global F0 level and dynamic range while retaining the shape of the source contour. For transforming F0 contours of lexical tones, the shapes of the source and target contour are distinct as shown in Fig. 1, and this linear method may not accurately transform F0 contours of the lexical tones.

2.4. GMM-based F0 Transformation

The statistical GMM-based approach is state-of-the-art F0 transformation [5], modeling the joint distribution of the source and target F0 contour by a GMM.

2.4.1. GMM-based F0 Transformation

Assume that the F0 value vector of the source and target speech are x and y respectively. State-of-the-art F0 transformation uses two-dimensional (2-D) F0 feature combined from static and delta F0 values. Aligned 2-D source and target F0 vectors x^{2D} are represented as in Eq. (3).

$$x^{2D} = \{X_i^T\}^T, \qquad y^{2D} = \{Y_i^T\}^T \tag{3}$$

where the combined 2-D F0 vector is given in Eqs. (4).

$$X_i = [x_i, \Delta x_i]^T, \qquad Y_i = [y_i, \Delta y_i]^T, \tag{4}$$

$i = 1..N$, N is the length of the aligned source and target vectors and ^{T} is the matrix transpose transform. Joint source-target 2-D vectors z can be represented as in Eq. (5).

$$z^{2D} = \{z_i\}, \qquad z_i = [{X_i}^T, {Y_i}^T]^T \tag{5}$$

The distribution of z is modeled by GMM λ, as in Eqs. (6) and (7).

$$p(z|\lambda) = \sum_{k=1}^{K} \alpha_k N(z; \mu_k, \Sigma_k), \tag{6}$$

$$\mu_k = \begin{bmatrix} \mu_k^x \\ \mu_k^y \end{bmatrix}, \Sigma_k = \begin{bmatrix} \sum_k^{xx} & \sum_k^{xy} \\ \sum_k^{yx} & \sum_k^{yy} \end{bmatrix} \tag{7}$$

where K is the number of Gaussian components.

$N(z; \mu_k, \Sigma_k)$ denotes the distribution with the mean μ_k and the covariance matrix Σ_k. α_k is the prior probability of z having been generated by component k. The parameters $(\alpha_k, \mu_k, \Sigma_k)$ can be estimated using the well-known expectation maximization (EM) algorithm.

In transformation phase, the suboptimum mixture component sequence $\hat{m}$ is determined by maximizing the likelihood $p(x|m, \lambda)$ as given in Eq. (8).

$$\hat{m} = \arg\max P(m|x, \lambda) \tag{8}$$

Having the suboptimum mixture component sequence, the converted F0 values $\hat{y}$ are determined by maximizing the likelihood $p(\hat{y}|x, m, \lambda)$ with respect to $\hat{y}$ as in [4].

2.4.2. Problems of GMM-based F0 Transformation for Converting Lexical Tones

In state-of-the-art GMM-based F0 transformation [5], transforming frame-based static F0 values and temporal deltas is efficient for expressive speech because F0 contour is largely varied on emotion of speech in short-time intervals. The experimental frame rate in previous researches [4, 5] is usually about 1 ms to 5 ms. Thus, the locality of frame-based F0 values is inside a short interval with duration about 1 ms to 5 ms. However, F0 contours of a source neutral unit and a target tonal unit are usually distinct in their approximate shapes rather than their details, as shown in Fig. 1. Therefore, transforming short-time frame-based F0 values and deltas may not improve the accuracy of transformed F0 contours of lexical tones but increase the noise sensitivity. In addition, general GMM-based approaches have many benefits but still have some drawbacks, including the insufficient precision of GMM models and parameters, the insufficient smoothness of the converted parameters between frames, and the over-smooth in each converted frame [8].

3. TD-GMM FRAMEWORK FOR SPECTRAL SEQUENCE TRANSFORMATION

To overcome general drawbacks of traditional GMM-based voice conversion (VC), presented in subsection 2.4.2, Nguyen .B and Akagi proposed TD-GMM framework in VC for spectral sequence transformation [8], with superior results compared with traditional GMM-based methods. In order to modify TD-GMM framework for F0 transformation, this section presents and investigates TD-GMM framework.

3.1. Temporal Decomposition

Altal [10] proposed TD as an interpolation method to decompose speech into a series of event functions and event target vectors, as given in Eq. (9).

$$\hat{y}(n) = \sum_{k=1}^{K} a_k \phi_k(n), 1 \leq n \leq N \tag{9}$$

where a_k and $\phi_k(n)$ are the k^{th} event target and event function, respectively. $\hat{y}(n)$ is the approximation of $y(n)$.

TD can be efficiently used in high-quality VC because it decomposes speech into mutual independent components event targets and event functions, and each component can be independently controlled and modified. However, the original TD is high computational complexity. To overcome this drawback, the modified restricted second order TD (MRTD) [11] with some improvements for event functions and targets estimations were proposed. MRTD has been considered as compact, flexible that is used in TD-GMM framework for VCs presented in the next sub-section.

3.2. TD-GMM Framework for Spectral Sequence Transformation

In TD-GMM framework, spectral parameters such as LSF parameters are decomposed into TD event targets and TD event functions by MRTD. Assume that there are M static TD event targets in the aligned source and target speech. Denote lsf_{xi}, lsf_{yi} are the TD event targets of LSF vectors (LSF-TD) of source speech x and target speech y, and $i = 1, 2, \cdots, M$. The 1-D source and target LSF-TD vectors lsf_x, lsf_y is represented as given in Eq. (10).

$$lsf_x = \{lsf_{xi}^T\}, \qquad lsf_y = \{lsf_{yi}^T\} \tag{10}$$

The joint source-target vector of event targets z is computed as in Eq. (11).

$$z = [lsf_x^T, lsf_y^T]^T \tag{11}$$

To improve the estimation of GMM parameters, the phoneme-based vector of TD event targets in [8] is also re-estimated by a normalization technique. The distribution of z is modeled by a GMM same as in conventional GMM methods, previously presented in Eqs. 6,7.

3.3. Problems when Adopting TD-GMM Framework to F0 Transformation

The result in [8] shows that converting static LSF-TD event targets can improve the estimation of GMM parameters as well as eliminate the frame-to-frame discontinuities in conventional GMM VCs, resulting natural and smooth converted speech. In general, TD-GMM framework might be applied for F0 contour also. However, it has been known that dynamic features of F0 are important but it can not be directly converted in TD-GMM framework.

In addition, the phoneme-based target alignment and training within TD-GMM framework [8] requires a huge data covering all phonemes to train all phoneme-based GMMs. This requirement might not be satisfied in this research when we attempt to accurately transform F0 contours of the lexical tones with limited data at offline stage, in order to improve the *QSr* of CSS.

4. PROPOSED METHODS FOR TRANSFORMING LEXICAL TONES

4.1. Modified TD-GMM Framework for Transforming F0 Contours of Lexical Tones

TD-GMM framework is efficient for transforming spectral sequence. Therefore, we attempt to adopt the original TD-GMM framework for transforming F0 contours of lexical tones. However, TD-GMM framework can not be directly and efficiently used in F0 transformation as presented in sub-section 3.3. Therefore, we modify TD-GMM framework for F0 transformation with the use of 2-D contextual F0-TD target vector, combined from static F0-TD targets and their 1^{st}-order deltas.

The use of contextual F0-TD target vector here is also to expand the temporal locality of frame-based F0 values and their deltas in state-of-the-art F0 transformation [5]. As presented in section 2.4.2, the locality of frame-based F0 values and their deltas is usually inside a short interval with duration about 1 ms to 5 ms. The locality of target-based F0 values and deltas here is within an interval with the duration about 40 ms with the experimental parameters in section 5.1. The locality of target-based F0 values and deltas can also be expanded longer by changing the the number of TD events per unit. It is shown in Fig. 1 that F0 contours of a source neutral unit and a target tonal unit are usually distinct in their approximate shapes rather than their details. Therefore, expanding the locality of the F0 values and deltas may improve the accuracy of the F0 transformation for lexical tones.

In addition, it has been known that low-dimensional vector is not suitable to be modeled by GMM because it might cause GMM over-fitting. Therefore, using delta features of F0 to extend the dimension of F0 vectors can improve the accuracy of the estimation of GMM parameters [5].

Assume that there are M static F0-TD targets the aligned source and target speech, $\{f0_i^x, f0^{y_i^t}\}$ are static F0-TD target sets of source F0 contour x and target F0 contour y^t, $i = 1, 2, \cdots, M$ with tone t^{th}, and $t = 2, 3, \cdots, T$. T is number of target tones, $T = 6$ in Vietnamese. The 2-D source and target F0-TD target vectors $f0^X, f0^{Y^t}$ is represented as given in Eqs. (12) , (13), and (14).

$$f0^X = [{f0_1^X}^T, ..., {f0_i^X}^T, ..., {f0_M^X}^T], \quad (12)$$

$$f0^{Y^t} = [{f0^{Y_1^t}}^T, ..., {f0^{Y_i^t}}^T, ..., {f0^{Y_M^t}}^T] \quad (13)$$

where

$$f0_i^X = [f0_i^x, \Delta f0_i^x]^T, \qquad f0^{Y_i^t} = [f0^{y_i^t}, \Delta f0^{y_i^t}]^T \quad (14)$$

The joint source-target vector of F0-TD targets z is computed as in Eq. (15).

$$z = [{f0^X}^T, {f0^{Y^t}}^T]^T \quad (15)$$

The distribution of z is modeled by a GMM same as in conventional GMM methods, previously presented in Eqs. (6) and (7), and the converted F0 contour with the tone t^{th} $\hat{y}^t$ are determined as in section 2.4.

4.2. Modified NNS-based Alignment for Transforming F0 Contour of Lexical Tones

The phoneme-based target alignment and training within TD-GMM framework for spectral sequence [8] can not be used if we have a limited training data as presented in subsection 3.3. The non-parallel alignment method using nearest neighbor search (NNS) [5] can be used with a limited training data. However, the alignment method in [5] searches the closest neighbors in the whole data space that may reduce the accuracy of the alignment.

In this research, we modify the NNS-based alignment in [5] integrated with the modified TD-GMM framework for F0 transformation by clustering available phonetic units into some clusters based on their articulatory similarities. Each cluster produces a phonetic-dependent subspace for searching in the modified NNS-based alignment. Thus, the source and target units for each aligned source-target pair are search from correspondent subspaces that the source/target units belong to.

When transforming F0 contour of lexical tones, spectral envelope parameters of all units in each same-phonation set are almost the same because they are related to similar vocal tract parameters caused by similar pronunciation behaviors. Thus, we use spectral envelope feature LSF for alignment instead of directly using F0. Then, we use the F0-TD targets at the positions of the aligned LSF-TD target pairs as inputs of phonetic-dependent GMM models for training.

In source-target alignment, assume that the source LSF-TD target vector computed from neutral units, is $\{lsf_m\}, m = 1, 2, \cdots, M$. When training for target tone t^{th}, $t = 2, 3, \cdots, T$ with $T = 6$ in Vietnamese, the set of all tonal units with tone t^{th} is $\hat{ws}^t$, and $\hat{ss}^{t,m}$ is a tonal subspace of $\hat{ws}^t$ containing all units belong to the phonetic unit cluster that lsf_m belongs to. The target vector for alignment is computed as in Eq. (16).

$$\tilde{lsf}_m = \text{NNS}(lsf_m, \hat{ss}^{t,m}), \hat{ss}^{t,m} \in \hat{ws}_t. \quad (16)$$

NNS function here returns the closest neighbors found in target space. The aligned LSF-TD target pairs therefore are $\{lsf_m, \tilde{lsf}_m\}$. To be used for F0 transformation, we need the positions of the aligned LSF-TD target pairs rather than their values. The positions of aligned pairs is $\{m, p(\tilde{lsf}_m)\}$ where $p(\tilde{lsf}_m)$ is the position of $\tilde{lsf}_m$.

We use also target-source alignment. Assume that the target LSF-TD target vector computed from tonal units with tone t^{th}, is $\{lsf_n^t\}$, source vector for alignment is computed as in Eq. (17).

$$\tilde{lsf}_n^t = \text{NNS}(lsf_n^t, \hat{ss}^{1,n}) \quad (17)$$

where $\hat{ss}^{1,n} \in \hat{ws}^1, n = 1, 2, \cdots, N$, $\hat{ws}^1$ is the set of all neutral units, $\hat{ss}^{1,n}$ is a neutral subspace of $\hat{ws}^1$ containing all neutral units belong to the phonetic unit cluster

that lsf_n^t belongs to. The positions of aligned pairs is $\{p(\tilde{lsf}_n^t), n\}$ where $p(\tilde{lsf}_n^t)$ is the position of $\tilde{lsf}_n^t$.

Combining both source-target and target-source alignments, the GMM transformation function F is trained from the aligned pairs of F0 vectors: $\{f0^X(m), f0^{Y^t}(p(\tilde{lsf}_m))\}$ and $\{f0^X(p(\tilde{lsf}_n^t)), f0^{Y^t}(n)\}$. Here, $f0^X, f0^{Y^t}$ are F0-TD target vectors combined from static F0-TD targets and their deltas of source neutral units and target tonal units with tone t^{th} respectively, same as in Eqs. (12) and (13).

5. EVALUATIONS

5.1. Data Preparation and Experimental Setup

Vietnamese is a typical tonal language [12] that we chose to evaluate the proposed F0 transformation for lexical tones. There are six distinct tones in Vietnamese as shown in Fig. 1. The corpus used for our evaluation is Vietnamese corpus DEMEN567 [7]. The total capacity of DEMEN567 corpus in WAV format is about 70 MB, the sampling rate is 11025 Hz and the resolution is 16 bits per sample. DEMEN567 includes 567 Vietnamese utterances, 7724 Vietnamese tonal syllables. Among these 7724 tonal syllables, there are 1088 distinct Vietnamese tonal syllables. In order to find out the best training condition for improving the quality and the *QSr*, the training dataset was scaled as shown in Table.1.

Each scaled dataset was divided into six groups, corresponding with six Vietnamese tones. The group of neutral units was used as the source while other five tonal unit groups were used as targets for F0 transformation. The numbers of syllables in each group were different between the tones. In each tonal unit group, 10 tonal syllables were used for testing, resulting total 50 tonal syllables were used for testing.

In Vietnamese, two kinds of syllable can be distinguished "open" and "closed" syllable. Closed syllables with the codas /p,/ /t/, /k/ could only be combined with tone rising and tone drop while open and other closed syllables could be combined with all six tones to become a meaningful tonal syllable. In our evaluation, we did not use "closed" syllables with final codas /p, t, k/. STRAIGHT was used to extract F0 and spectral envelope (SE), same as in [8]. The frame size and update interval were set to 20 ms and 1 ms respectively in both experiments. The GMM-based F0 transformations used LSF, computed from SE of STRAIGHT, with the order P set to be 32 for the alignments. The number of GMM mixtures was 4. When using TD analysis/synthesis, each phoneme was represented by five F0-TD targets.

5.2. Evaluations on Speech Quality and QSr

The MOS test was used to evaluate the quality of the tonal syllables transformed by three comparative methods, including the conventional F0 transformation transforming the means and variances of F0 [4], the state-of-the-art F0 transformation using frame-based GMM [5], and the proposed F0 transformation. The MOS test was also conducted with original tonal syllables and tonal syllables transformed by the F0 exchange method in [6] that used full Vietnamese tonal syllables dataset with capacity about 60 MBs. In all MOS tests, the subjects were five native Vietnamese subjects who were asked to rate the perceptual quality of the converted tonal syllables on a five-point scale (1: bad, 2: poor, 3: fair, 4: good and 5: excellent).

QSr was mentioned as a target in researches of Kain et. al [2, 3]. However, they did not define clearly this ratio. In this paper, we defined the *QSr* as in Eqs. (18) and (19).

$$QSr = \bar{S}_{\text{MOS}}/\bar{S}_{\text{Size}} \tag{18}$$

$$\bar{S}_{\text{MOS}} = S_{\text{MOS}}/5, \bar{S}_{\text{Size}} = S_{\text{Size}}/\lim(S_{\text{Size}}). \tag{19}$$

where S_{MOS} is MOS score, S_{Size} is the database size, $\lim(S_{\text{Size}})$ is the upper limitation of the database size.

The simple tone transformation by exchanging F0 contours [6] requires a full data covering all tonal units to compute all tone parameters at offline stage. In Vietnamese, there are 5489 non-neutral tonal syllables. Using the same recording parameters as in the corpus DEMEN567, the capacity of the whole tonal units dataset is about 60 MBs. We used this size as the upper limitation in Eq. (19).

The results on MOS scores are shown in Fig. 2. The MOS scores of the three first methods (training-based methods) depend on the scaled datasets while those of the two later ones are independent with the scaled datasets. However, we draw all results in one unique figure to easily compare each method with other ones. This result shows that our proposed TD-GMM F0 transformation outperformed both conventional [4] and state-of-the-art [5] F0 transformations in terms of speech quality. The quality of tonal syllables converted by proposed F0 transformation using the training dataset with the capacity about 30 MBs (dataset 5^{th}) was equivalent with the quality of the F0 exchange method in [6] using offline dataset with capacity about 60 MBs.

The results on *QSr* are shown in Fig. 3. The *QSr* of the proposed method (a training-based method) depends on the scaled datasets while that of F0 exchange method is independent with the scaled datasets. However, we also draw all results in one unique figure to easily compare. This result shows that the proposed method trained with all scaled datasets (capacity from about 10 MBs to 30 MBs) improved the *QSr* compared with the F0 exchange method in [6] using offline dataset with capacity about 60 MBs.

Table 1. Scaled datasets for training

Scaled Dataset Index	No. of Syllables	Size (MBs)
1	500	9.8
2	1000	16.4
3	1500	19.6
4	2000	24.3
5	2500	27.2

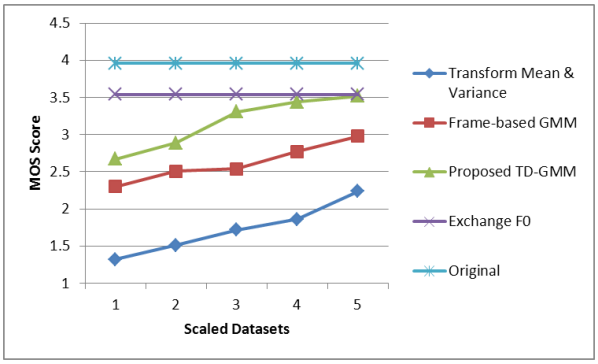

Fig. 2. MOS scores for F0 transformations.

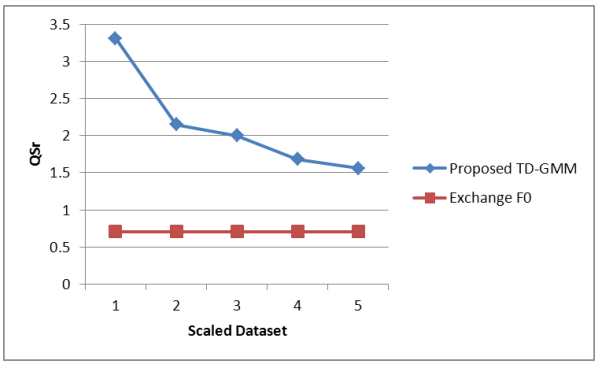

Fig. 3. QSr with F0 exchange and the proposed F0 transformation.

6. CONCLUSIONS

In this paper, we proposed an F0 transformation for lexical tones using a modified TD-GMM framework with a modified NNS non-parallel alignment. The experimental results show that our proposed F0 contour transformation outperformed conventional and state-of-the-art F0 transformations for transforming lexical tones in terms of speech quality. The evaluation results also show that our proposed method improves the *QSr* compared with simple F0 exchange method. As a consequence, the proposed method is efficient in both footprint reduction and availability with limited data while preserves perceptual quality. Therefore, the proposed method can be efficiently applied for CSS of tonal languages.

7. ACKNOWLEDGEMENTS

This study was supported by A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).

8. REFERENCES

[1] Shoham, T., Malah, D., and Shechtman, S., "Quality Preserving Compression of a Concatenative Text-To-Speech Acoustic Database,"*IEEE Trans. on Audio, Speech, and Language Proc.,* Vol. 20, No. 3, pp. 1056–1068, (2012).

[2] Kain, A., Miao, Q., and van Santen, J., "Spectral control in concatenative speech synthesis," *Proc. ISCA Workshop on Speech Synthesis,* (2007).

[3] Kain, A. and Leen, T., "Compression of Line Spectral Frequency Parameters using the Asynchronous Interpolation Model," *Proc. of 7th ISCA Workshop on Speech Synthesis,* (2010).

[4] Toda, T., Black, A., Tokuda, K., "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Language Process,* v15-i8, pp. 2222-2235, (2007).

[5] Wu, Z., Kinnunen, T. Chng, E.S, Li, H., "Text-Independent F0 Transformation with Non-Parallel Data for Voice Conversion," *Proc. Interspeech 2010,* pp. 1732-1735, (2010).

[6] Luksaneeyanawin, S., Tone transformation, *Proc. SNLP-95*, pp. 345-360, (1995).

[7] Dung, T.N., Mixdorff, H., "Fujisaki Model based F0 contours in Vietnamese TTS," *Proc. ICSLP-04*, (2004).

[8] Nguyen, B. and Akagi, M., "Efficient modeling of temporal structure of speech for applications in voice transformation," *Proc. Interspeech 2009,* pp. 1631-1634, (2009).

[9] Nguyen, VL., Edmondson, A., "Tones and voice quality in modern northern Vietnamese: Instrumental case studies," *Mon-Khmer Studies* 28: pp. 1-18, (1998).

[10] Atal, B.S, "Efficient coding of LPC parameters by temporal decomposition," *Proc. ICASSP 1983,* pp. 81-84, (1983).

[11] Nguyen, C., Ochi, T., and Akagi, M., "Modified restricted temporal decomposition and its application to low bit rate speech coding," *IEICE Trans. on Inf. and Sys.,* vol. E86-D, pp. 397-405, (2003).

[12] Hoang, P., "Vietnamese Grammar (Vietnamese)," *Da nang Publisher*, pp. 9-15, (2003).