INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Noise robust pitch stylization using minimum mean absolute error criterion

Chiranjeevi Yarra[1], *Prasanta Kumar Ghosh* [2]

[1] Language Technologies Research Center (LTRC), IIIT Hyderabad, 500032, India
[2] Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India

[1]chiranjeevi.yarra@iiit.ac.in, [2]prasantg@iisc.ac.in

Abstract

We propose a pitch stylization technique in the presence of pitch halving and doubling errors. The technique uses an optimization criterion based on a minimum mean absolute error to make the stylization robust to such pitch estimation errors, particularly under noisy conditions. We obtain segments for the stylization automatically using dynamic programming. Experiments are performed at the frame level and the syllable level. At the frame level, the closeness of stylized pitch is analyzed with the ground truth pitch, which is obtained using a laryngograph signal, considering root mean square error (RMSE) measure. At the syllable level, the effectiveness of perceptual relevant embeddings in the stylized pitch is analyzed by estimating syllabic tones and comparing those with manual tone markings using the Levenshtein distance measure. The proposed approach performs better than a minimum mean squared error criterion based pitch stylization scheme at the frame level and a knowledge-based tone estimation scheme at the syllable level under clean and 20dB, 10dB and 0dB SNR conditions with five noises and four pitch estimation techniques. Among all the combinations of SNR, noise and pitch estimation techniques, the highest absolute RMSE and mean distance improvements are found to be 6.49Hz and 0.23, respectively.

Index Terms: Pitch stylization, minimum MAE criterion, dynamic programming based segmentation, noise robustness

1. Introduction

Pitch stylization is a process of representing the pitch contour compactly with the least number of segments [1]. The stylized pitch contour has been shown to be useful in several applications including emotion recognition [2, 3], computer assisted language learning (CALL) [4, 5], speech synthesis [3, 6], phrase boundary detection [7], disfluency identification, speaker verification [8] and intonation modeling [9, 10]. Origilia et al. have emphasized that the stylized pitch should be robust to unimportant variations, otherwise those variations would introduce noise in the applications [11].

To remove these unwanted variations, Rossi et al. have applied a low-pass filter on the pitch contour during the pitch stylization process [12]. Instead, Demenko et al. have performed pitch stylization directly on the pitch contour considering the segments obtained with manual markings [5]. However, manual syllable segmentation is cumbersome. To avoid manual intervention, Uwe et al. have performed pitch stylization by obtaining segments automatically and applying a line-fit on each segment separately [13]. But, in this two-step approach, the errors in the segmentation affects the stylization process. On the other hand, Ghosh et al. have proposed dynamic programming (DP) based approach to obtain the segments and stylized pitch jointly [14]. Origlia et al. have shown that the DP based approach is not statistically different from the other approaches that use syllable segments in a subjective manner [15]. Most of these works consider the stylization process by minimizing mean squared error (MSE) between a pitch contour and its stylized contour [13, 5, 16, 14].

However, an estimated pitch contour often suffers from halving and doubling errors. For noisy speech, these errors are more than those under clean conditions [17]. Though these errors occur naturally under creaky voice conditions [18], such phenomena are less often, hence not considered in this work. Figure 1 illustrates the errors (halving/doubling) and their effect on pitch stylization with an exemplary voiced segment. It shows the ground truth pitch (green colored) as well as estimated pitch (black colored) obtained using sub-harmonic to harmonic ratio (SHR) [19] under clean (Figure 1a) and additive white Gaussian noise at 0dB SNR condition (Figure 1b). The figure shows that the errors in the estimated pitch (blue rectangular boxes) are more under 0dB SNR compared to those in clean condition. Thus, the stylization approach has to be robust in the presence of such errors. However, from the figure, the stylized pitch (red coloured) obtained using typical MSE based criterion on the estimated pitch is not close to the ground truth pitch (higher MSE) under 0dB SNR compared to that under clean condition. This suggests that the MSE based stylization could be erroneous in pitch halving and doubling errors as it is sensitive to outliers caused by the errors.

In the literature, it has been shown that the outliers can be handled effectively using a minimization criterion with mean absolute error (MAE) than that with MSE [20]. However, it is challenging to formulate MAE based criterion for the pitch stylization since it requires joint optimization to obtain segment boundaries and perform stylization on each of these segments. In this work, we address this challenge by proposing a DP approach with the MAE cost function. We show that this formulation is robust under four pitch estimation techniques, five noise types and three SNR conditions (20dB, 10dB and 0dB) at the frame level closeness and the syllable level tone embeddings considering three corpora, namely, KEELE [21], PaulBaghsaw [22] and British English (BE) training corpus [23]. We compare the stylized pitch closeness with the ground truth pitch extracted from a laryngograph signal by computing root mean squared error (RMSE) in a voiced segment averaged across all segments in a corpus. We measure the tone embeddings' quality by computing Levenshtein distance between estimated tones from the stylized pitch and manual tone markings in an utterance averaged across all utterances in a corpus. The averaged RMSE and distance are less with the proposed pitch stylization than that obtained respectively from the MSE-based baseline pitch stylization scheme and knowledge-based baseline tone estimation scheme under all the noise, SNR and pitch estimation technique combinations.

2. Database

We use KEELE [21] and PaulBaghsaw (PB) [22] corpora for all experiments at the frame level in this work. KEELE database consists of utterances from five male, five female and five children speakers. PB database consists of 50 sentences spoken by one male and one female speaker. In the experiments, we

http://dx.doi.org/10.21437/Interspeech.2021-1307

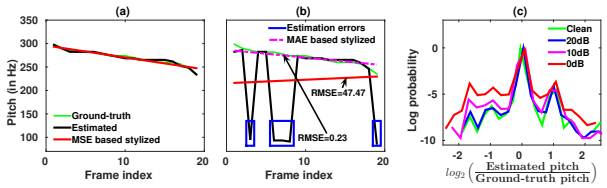

Figure 1: *Illustration shows the motivation for MAE based criterion, where the estimated pitch is obtained using SHR method under a) clean condition and b) additive white Gaussian noise at 0dB SNR condition c) pitch estimation error patterns using the distribution of* $log_2\left(\frac{Estimated\ pitch}{Ground\ truth\ pitch}\right)$.

consider only the sentences of all male and female subjects from corpora. In both the corpora, each spoken utterance has been recorded simultaneously with a laryngograph signal used to compute the reference pitch and considered it as the ground truth. Further, for the experiments in syllable level, the speech data is considered from a spoken English training material used for teaching BE, referred to as BE training corpus [23]. The speech recordings selected for our experiments contain all the utterances of intonation phrases belonging to intonation lessons. The entire speech recording is manually segmented into individual speech files belonging to every utterance and obtained tone sequence containing four tones – rise, fall, low and high for each utterance [24]. The entire speech data contains a total of 233 utterances and has been spoken by one male and one female native BE speaker. Also, in the experiments, we use five noises: babble, f16, hfc, volvo and white from NOISEX-92 database [25].

3. Proposed approach

3.1. Preliminaries

Let $\{x_n\}_{n=1}^N$ be a pitch contour representing a set of N discrete values where n is an index variable. The problem of pitch stylization is to approximate the $\{x_n\}_{n=1}^N$ with K piecewise polynomials to obtain a stylized pitch contour $\{\hat{x}_n\}_{n=1}^N$. Let $\hat{x}_n$ in the k-th segment having begin and end indices $\lambda_1(k)$ and $\lambda_2(k)$ is represented with P-th order polynomial as $\sum_{p=0}^{P} \alpha_p(k) n^p$, where $\alpha_p(k), 0 \leq p \leq P$ are the polynomial coefficients. Further, in order to ensure the continuity across the segment boundaries, it is assumed that $\lambda_1(k+1) = \lambda_2(k)$, $1 \leq k \leq N-1$, $\lambda_1(1) = 1$ and $\lambda_2(K) = N$. In addition, we assume that the pitch values are realizations of the random variables X at index n, denoted by $X_n = \hat{x}_n + \eta_n$, where η_n denotes the noise, which is independent identically distributed (i.i.d.) random variable. In the existing work [14], η_n has been assumed as white Gaussian noise with mean 0 and variance 1. Under this assumption, it is trivial to show that the maximum likelihood (ML) solution of the $\hat{x}_n$ can be obtained as $\underset{\{\hat{x}_n\}}{\operatorname{argmin}} \sum_{n=1}^{N} (x_n - \hat{x}_n)^2$. Replacing $\hat{x}_n$ in the k- th segment with $\sum_{p=0}^{P} \alpha_p(k) n^p$, the optimization problem becomes

$$\underset{\{\lambda_1\},\{\lambda_2\},\{\alpha_p\}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{n=\lambda_1(k)}^{\lambda_2(k)} \left(x_n - \sum_{p=0}^{P} \alpha_p(k) n^p \right)^2 \quad (1)$$

The parameters $\{\lambda_1\}, \{\lambda_2\}$ and $\{\alpha_p\}$ are solved using DP [14]. In this work, we replace MSE with MAE criterion.

3.2. Motivation for MAE criterion

In order to examine the need for MAE criterion, we analyze the halving and doubling errors in pitch estimation using SHR in the KEELE corpus under additive white Gaussian noise conditions (Figure 1c). The figure shows the distribution of logarithm of the ratio (LR) between the estimated pitch and the ground-truth pitch $\left(\text{LR} = log_2\left(\frac{\text{Estimated pitch}}{\text{Ground truth pitch}}\right)\right)$ under clean as well as 20dB, 10dB and 0dB SNR conditions. Zero, 1, -1 values of LR indicate 0, halving and doubling errors in the estimated pitch. From the figure, it is observed that, apart from LR=0, there are significant peaks at LR=-1, 1, 1.585 under clean and all three SNR conditions. These peaks cause the tail of the distribution to be heavy, indicating that the error distribution may not be Gaussian, which corresponds to the MSE criterion based pitch stylization. Instead, we assume the error distribution to be Laplacian.

3.3. MAE based approximation

It is trivial to show that the ML solution of $\hat{x}_n$ for minimizing MAE is obtained as $\underset{\{\hat{x}_n\}}{\operatorname{argmin}} \sum_{n=1}^{N} |x_n - \hat{x}_n|$, when η_n is assumed to be an i.i.d. Laplacian noise with location and scale parameters as zero and one respectively. Now, for k-th segment, replacing $\hat{x}_n$ with $\sum_{p=0}^{P} \alpha_p(k) n^p$, the optimization problem becomes:

$$\underset{\{\lambda_1\},\{\lambda_2\},\{\alpha_p\}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{n=\lambda_1(k)}^{\lambda_2(k)} \left| x_n - \sum_{p=0}^{P} \alpha_p(k) n^p \right| \quad (2)$$

For this equation, first, we derive the steps to find optimal parameters for $K = 1$ and analyse its effectiveness in the pitch stylization with respect to MSE based criterion with $K = 1$ in (1). Later, we derive the steps to find optimal parameters for any value of K.

When $K = 1$, $\lambda_1(K) = s = 1$ and $\lambda_2(K) = r = N$. Considering these, (2) becomes $\underset{\{\alpha_p\}}{\operatorname{argmin}} \sum_{n=s}^{r} |x_n - \sum_{p=0}^{P} \alpha_p n^p|$. Further, it can be represented in matrix-vector form as follows:

$$\underline{\hat{\alpha}}(s,r) = \underset{\underline{\alpha}}{\operatorname{argmin}} |A\underline{\alpha} - \underline{x}|;\ E(s,r) = \min_{\underline{\alpha}} |A\underline{\alpha} - \underline{x}| \text{ where}$$

$$A = \begin{bmatrix} s^0 & \dots & s^P \\ (s+1)^0 & \dots & (s+1)^P \\ \vdots & \ddots & \vdots \\ r^0 & \dots & r^P \end{bmatrix},\ \underline{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_P \end{bmatrix},\ \underline{x} = \begin{bmatrix} x_s \\ x_{s+1} \\ \vdots \\ x_r \end{bmatrix} \quad (3)$$

Solving for $\underline{\alpha}$ using (3) is identical to solving the $\underline{\alpha}$ from the following: $\underset{\underline{\alpha},\underline{\theta}}{\operatorname{argmin}} \sum_{n=1}^{N} \theta_n$ subject to $A\underline{\alpha} - \underline{x} \leq \underline{\theta}$ and $A\underline{\alpha} - \underline{x} \geq -\underline{\theta}$, where $\underline{\theta} = [\theta_1, \theta_2, \dots, \theta_N]^T$. This can further be written as

$$\underline{\hat{\phi}} = \underset{\underline{\phi}}{\operatorname{argmin}}\ \underline{f}^T \underline{\phi} \quad \text{subject to } D\underline{\phi} - \underline{y} \leq 0 \quad (4)$$

where $D = \begin{bmatrix} A & -I \\ -A & -I \end{bmatrix}$, $\underline{y} = \left[\underline{x}^T, -\underline{x}^T\right]^T$, $\underline{\phi} = \left[\underline{\alpha}^T, \underline{\theta}^T\right]^T$ and $\underline{f} = [\underline{0}, \underline{1}]$, where $I_{N \times N}$ is an identity matrix and $\underline{0}$ and $\underline{1}$ are the vectors of length $P+1$ and N containing zeros and ones respectively. We propose to solve (4) for $\underline{\phi}$ using linear programming approach, from which $\underline{\alpha}$ is obtained.

The dotted magenta color line in Figure 1b shows the stylized pitch obtained using (2) with K=1 for the exemplary pitch segment shown in Figure 1. From the figure, it is observed that the MAE based stylized pitch is closer to the ground truth pitch

compared to the MSE based stylized pitch. Further, the root mean squared error (RMSE) between the stylized and ground truth pitch is found to be lower when (2) is used compared to that when (1) is used for stylization. This indicates that the stylized pitch obtained with MAE based criterion is more robust to the doubling and halving errors compared to those estimated with the MSE based criterion.

3.4. MAE based piecewise approximation

For any value of $K \geq 1$, it is required to find the best polynomial fit and the boundary points for each segment. Ghosh et al. [14] proposed an approach to find the best polynomial and boundary points using a DP based approach. Their approach works on the principle of locally best polynomial fit within the k-th segment to achieve a globally best solution. Following their work, we perform pitch stylization considering MAE in (2) as follows:

1) Local best solution: For given local data points $\{x_n\}_{n=\lambda_1(k)}^{\lambda_2(k)}$ of the k-th segment, the best polynomial approximation is obtained subject to the constraint of $\sum_{p=0}^{P} \alpha_p(k)\lambda_1(k)^p = \sum_{p=0}^{P} \alpha_p(k-1)\lambda_1(k)^p$. In a matrix-vector form, this constraint can be represented as $h\underline{\phi} = b\left(\underline{\alpha}(k-1), \lambda_1(k)\right)$, where $b\left(\underline{\alpha}(k-1), \lambda_1(k)\right) = \sum_{p=0}^{P} \alpha_p(k-1)\lambda_1(k)^p$, $h = [\lambda_1(k)^0, \lambda_1(k)^1, \ldots, \lambda_1(k)^P, \underline{0}]$ and $\underline{0}$ is an all zero vector of length $\lambda_2(k) - \lambda_1(k) + 1$. In order to obtain the polynomial under this constraint based on MAE, we propose to solve (4) by including the constraint $h\underline{\phi} = b\left(\underline{\alpha}(k-1), \lambda_1(k)\right)$ to the existing set of constraints. This also involves computing A and $\underline{x}$ considering $s = \lambda_1(k)$ and $r = \lambda_2(k)$ and taking length of $\underline{1}$ in $\underline{f}$ as $\lambda_2(k) - \lambda_1(k) + 1$. The vector containing the optimal polynomial coefficients and respective MAE are referred to as $\hat{\underline{\alpha}}\left(\lambda_1(k), \lambda_2(k), b\right)$ and $\mathrm{E}\left(\lambda_1(k), \lambda_2(k), b\right)$ respectively.

2) Global best solution: Considering the local best solution $\hat{\underline{\alpha}}\left(\lambda_1(k), \lambda_2(k), b\right)$ and $\mathrm{E}\left(\lambda_1(k), \lambda_2(k), b\right)$, we obtain the optimal values $\lambda_1^*(k), \lambda_2^*(k), \hat{\underline{\alpha}}^*(k)\ \forall 1 \leq k \leq K$ using DP as described in Algorithm 1.

Algorithm 1 MAE based piece-wise polynomial. Input: K, P, $\{x_n\}_{n=1}^{N}$ and output: $\lambda_1^*(k), \lambda_2^*(k), \hat{\underline{\alpha}}^*(k)\ \forall 1 \leq k \leq K$

Initialization: Compute $e_1(r) = E(1, r)$ and $\gamma_1(r) = \hat{\underline{\alpha}}(1, r)$ using (3); $b_1(r) = b(\gamma_1(r), r)$; $\xi_1(r) = 1\ \forall\ P+1 \leq r \leq N$

Forward-pass:

for $2 \leq k \leq K$ & $kP + 1 \leq r \leq N$ **do**

Compute $b_k(r) = b(\gamma_{k-1}(r), r)$

Compute $\mathrm{E}(s, r, b_k(r))$ using (4) with additional constraint of $h\underline{\phi} = b_k(r)\ \forall\ 1 \leq s \leq r - P$

$e_k(r) = \min_{1 \leq s \leq r-P} \{e_{k-1}(r) + \mathrm{E}(s, r, b_k(r))\}$

$\xi_k(r) = \operatorname{argmin}_{1 \leq s \leq r-P} \{e_{k-1}(r) + \mathrm{E}(s, r, b_k(r))\},\ \gamma_k(r) = \hat{\underline{\alpha}}\left(\xi_k(r), r, b_k(r)\right)$

end for

Back-track:

$\lambda_2^*(K) = N;\ \lambda_1^*(K) = \xi_K(N)$

for k from K to 2 **do**

$s = \lambda_1^*(k),\ r = \lambda_2^*(k),\ \hat{\underline{\alpha}}^*(k) = \hat{\underline{\alpha}}(s, r, b_k(s)),\ \lambda_2^*(k-1) = s,\ \lambda_1^*(k-1) = \xi_{k-1}(s)$

end for

$\hat{\underline{\alpha}}^*(1) = \hat{\underline{\alpha}}\left(\lambda_1^*(1), \lambda_2^*(1)\right)$

4. Experiments and results

4.1. Experimental set-up

For the experiments, we add noise to each speech signal at SNRs of 20dB, 10dB and 0dB. We obtain stylized pitch considering the pitch estimated under clean and all three SNR conditions using four pitch estimation techniques, namely, SHR, SWIPE [26], PEFAC [27] and YIN [28]. To determine the value of K for each voiced segment, we follow an approach similar to that proposed by Wang et al. [16], where a Wavelet decomposition of the pitch contour is performed using Daubechies wavelet (Db10), and the number of extrema in level 3 of the decomposition is used as $K - 1$. The voiced segments are obtained using ground truth pitch and used in experiments under all noise and SNR conditions. We analyze the proposed approach performance in two folds: 1) frame level stylized pitch values' accuracy, and 2) the effectiveness of stylized pitch in preserving perceptually relevant information at the syllable level, i.e., syllabic tones' accuracy.

4.1.1. Set-up for frame level:

We compute RMSE between the ground truth pitch and stylized pitch in each voiced segment and consider its mean across all voiced segments in each corpus as the objective measure, referred to as mean RMSE. We consider the work proposed by Ghosh et al. [14], which MSE based criterion, as the baseline. For this experiment, we consider KEELE and PB corpora.

4.1.2. Set-up for syllable level:

We compute Levenshtein distance between manual (the ground truth) tone sequence and predicted tone sequence from the stylized pitch in each utterance and consider its mean across all the utterances in BE training corpus as the objective measure, referred to as mean distance. We follow the work proposed by Mertens to predict the tone sequence from the stylized pitch obtained in the proposed approach [29]. We consider tone sequence estimated with the Prosogram tool [30] from the estimated pitch as the baseline. In both the tone sequence estimation, we use the Glissando threshold as $\frac{16}{T}$, where T is the time duration of the segment.

Table 1: *Averaged improvements (in Hz) across all five noise types in mean RMSE for both the corpora using polynomial order $P = 1$ and $P = 2$ under clean and all three SNRs for all four pitch estimation techniques.*

		$P = 1$				$P = 2$			
		Clean	20dB	10dB	0dB	Clean	20dB	10dB	0dB
KEELE	SHR	0.42	0.60	1.78	4.81	0.25	0.33	1.30	4.18
	SWIPE	0.17	0.39	1.07	3.25	0.46	0.48	0.95	2.78
	PEFAC	1.70	1.52	1.97	2.14	1.16	1.23	1.44	1.65
	YIN	1.11	0.83	2.38	3.72	0.72	0.78	1.96	2.62
	Avg	0.82	0.84	1.80	3.48	0.65	0.70	1.41	2.81
PB	SHR	1.14	1.31	2.71	6.49	1.11	1.28	2.43	4.93
	SWIPE	0.38	0.52	1.19	3.08	0.18	0.27	0.58	1.78
	PEFAC	1.28	1.09	1.13	1.73	0.61	0.61	0.86	1.33
	YIN	0.65	0.66	1.24	3.65	0.72	0.74	1.13	2.67
	Avg	0.86	0.90	1.57	3.74	0.66	0.73	1.25	2.68

4.2. Results and discussions

4.2.1. Frame level

Table 1 shows the improvement with the proposed MAE criterion in mean RMSE over the baseline MSE criterion for both the corpora averaged across all five noise types under clean and all

three SNR conditions for $P = 1$ and $P = 2$ considering all four pitch estimation techniques. From the table, it is observed that all the entries in the table are positive. This indicates that the pitch stylization errors are reduced with the proposed method compared to the baseline method. When we average the improvements across all five noise types and all four pitch estimation techniques (indicated in the blue color), it is observed that the averaged improvements monotonically increase from clean to 0dB SNR in both the corpora for both P=1 and P=2. Even the monotonic increment in the improvements is consistently observed in all four pitch estimation techniques as well. These together indicate that, on average, the proposed MAE based pitch stylization performs better than the MSE based baseline with a larger margin at lower SNRs.

4.2.2. *Syllable level*

Table 2 shows the averaged mean distances across all five noise types under clean and all three SNR conditions for each pitch estimation method. The table shows that the averaged mean distance obtained in four pitch estimation methods and their average are lower than that obtained using the baseline Prosogram tool under clear and all three SNR conditions. This indicates the benefit of the proposed pitch stylization in preserving perceptual relevant information in terms of the tones. Further, from the table, it is observed that the difference in mean distances between the proposed approach and Prosogram is the least under clean condition and the highest in 0dB SNR condition under all five noise types. This suggests that the proposed stylization robust to the pitch estimation inaccuracies caused by the noise in the tone estimation.

Table 2: *Mean distances and averaged mean distances across all five noise types for BE training corpus considering $P = 1$ under clean and all three SNRs for all four pitch estimation techniques.*

	SHR	SWIPE	PEFAC	YIN	Avg	Prosogram
Clean	0.62	0.60	0.66	0.58	0.61	0.83
20dB	0.62	0.62	0.66	0.58	0.62	0.83
10dB	0.63	0.63	0.67	0.60	0.63	0.85
0dB	0.66	0.66	0.69	0.64	0.67	0.90

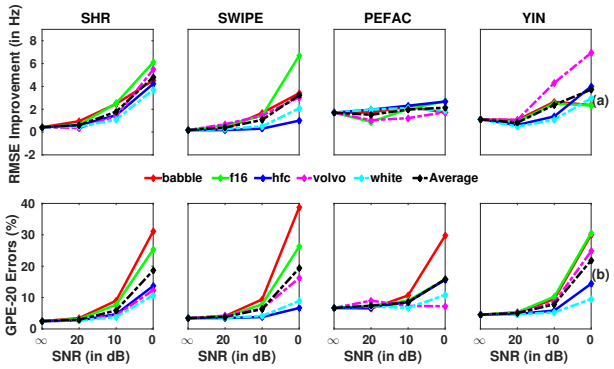

Figure 2: *(a) SNR specific variations in the RMSE improvements (in Hz) and (b) GPE-20 errors (%) for all five noise types and all four pitch estimation techniques on KEELE corpora.*

4.2.3. *Analysis*

Further, we analyze the improvements in each noise condition with the proposed approach on KEELE corpora at clean and all three SNRs for $P = 1$ using Figure 2a. To examine how these improvements depend on the pitch estimation errors, we compute gross pitch estimation (GPE)-20 error and plot the same for each noise under clean and all three SNR conditions for all four pitch estimation techniques in Figure 2b. The GPE-20 error is computed as $100 \times N_{err}/N_v$, where, N_{err} is the total number of voiced frames, in which the estimated pitch values fall outside $\pm 20\%$ of the ground-truth pitch value and N_v is the total number of voiced frames. In both the figures, the black color line indicates averaged improvements and GPE-20 errors across all noises. The figure shows that the averaged improvements increase with an increase in the averaged GPE-20 errors. This is also true in most noise and SNR combinations for all four pitch estimation techniques. These together suggest that the proposed method is robust to the typical pitch estimation errors.

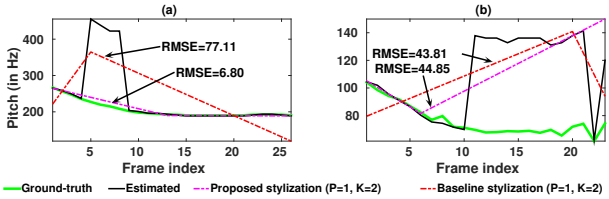

Figure 3: *Examplary voiced segments illustrating the merits and demerits of the proposed pitch stylization method. Pitch in these examples are estimated using SHR.*

We further investigate the merits and demerits of the proposed approach using Figure 3 with two exemplary voiced segments taken from KEELE database. Figure 3a and 3b show the stylized pitch contours corresponding to the voiced segments for which the RMSE is higher and lower respectively with the baseline compared to that with the proposed approach. From Figure 3a, it is observed that the stylized pitch with the proposed approach follows the trend in the ground truth pitch, although there is pitch estimation error. However, the stylized pitch using the baseline completely misses the trend in the ground-truth pitch. This could be because the baseline is more sensitive to the sudden pitch transitions. On the other hand, the baseline results in a lower RMSE in Figure 3b compared to that with the proposed stylization, where both the stylized pitch contours miss the trend in the ground truth pitch due to error in the pitch estimation over a relatively longer duration. However, the stylized pitch contour from the baseline follows the sudden transition that occurred at 23-rd frame at which the estimated pitch value is correct, which, in turn, results in a lower RMSE. This suggests that the proposed stylization could miss transitions, if present, at a shorter duration in the ground truth pitch.

5. Conclusions

Pitch stylization is performed by fitting an optimal piece-wise polynominal of order P. The entire pitch contour is divided into K segments using DP by minimizing MAE between stylized pitch and the estimated pitch. We found that the MAE based pitch stylization is more robust to the typical pitch estimation errors. Experiments with KEELE, PaulBaghsaw and BE training corpora reveal that the proposed approach performs better than the MSE based baseline and Prosogram baseline for all five noises under clean and 20dB, 10dB and 0dB SNR conditions considering four pitch estimation techniques. Further investigations are required to develop a method that incorporates complementary information from the proposed and baseline strategies for suppressing erroneous transitions. Future work also includes exploiting the MAE based criterion on explicitly obtained syllable boundaries.

6. References

[1] J. Hart and J. de Pijper, *Experiments on the Stylization of British English Intonation*. De Gruyter Mouton, 2019, pp. 570–573.

[2] C. Montacié and M.-J. Caraty, "Vocalic, lexical and prosodic cues for the interspeech 2018 self-assessed affect challenge." in *Interspeech*, 2018, pp. 541–545.

[3] S. Ma, D. Mcduff, and Y. Song, "Neural TTS stylization with adversarial and collaborative games," in *International Conference on Learning Representations*, 2019, pp. 1–16.

[4] C. Yarra and P. K. Ghosh, "An automatic classification of intonation using temporal structure in utterance-level pitch patterns for British English speech," in *IEEE India Council International Conference (INDICON)*, 2018, pp. 1–6.

[5] G. Demenko, A. Wagner, N. Cylwik, and O. Jokisch, "An audiovisual feedback system for acquiring L2 pronunciation and L2 prosody," *International Workshop on Speech and Language Technology in Education (SLATE)*, pp. 113–116, 2009.

[6] J. A. Louw and A. Moodley, "Automatic stylization, coding and modelling of intonation in text-to-speech for under-resourced languages," in *IEEE Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, 2017, pp. 197–202.

[7] A. Lacheret-Dujour, N. Obin *et al.*, "Automatic modelling and labelling of speech prosody: What's new with SLAM+?" in *International Congress of Phonetic Sciences (ICPhS)*, 2019.

[8] L. Mary, "Modeling and fusion of prosody for speaker, language, emotion, and speech recognition," in *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*. Springer, 2019, pp. 45–56.

[9] A. E. Rodgers, "K-Max: a tool for estimating, analysing, and evaluating tonal targets," in *International Conference on Speech Prosody*, 2020, pp. 225–229.

[10] P. Mertens, *From pitch stylization to automatic tonal annotation of speech corpora*. John Benjamins; Amsterdam, 2019.

[11] A. Origlia, G. Abete, F. Cutugno, I. Alfano, R. Savy, and B. Ludusan, "A divide et impera algorithm for optimal pitch stylization," *Interspeech*, pp. 1993–1996, 2011.

[12] P. S. Rossi, F. Palmieri, and F. Cutugno, "A method for automatic extraction of Fujisaki-model parameters," *in International Conference on Speech Prosody*, pp. 615–618, 2002.

[13] U. D. Reichel and K. Mády, "Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian," *in Interspeech*, pp. 111–115, 2014.

[14] P. K. Ghosh and S. S. Narayanan, "Pitch contour stylization using an optimal piecewise polynomial approximation," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 810–813, 2009.

[15] A. Origlia, G. Abete, and F. Cutugno, "A dynamic tonal perception model for optimal pitch stylization," *Computer Speech & Language*, vol. 27, no. 1, pp. 190–208, 2013.

[16] D. Wang and S. Narayanan, "Piecewise linear stylization of pitch via wavelet analysis," *in European Conference on Speech Communication and Technology*, pp. 3277–3280, 2005.

[17] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "Adaptive refinements of pitch tracking and HNR estimation within a vocoder for statistical parametric speech synthesis," *Applied Sciences*, vol. 9, no. 12, pp. 1–23, 2019.

[18] T. G. Csapó, G. Németh, and M. Cernak, "Residual-based excitation with continuous F0 modeling in HMM-based speech synthesis," in *International Conference on Statistical Language and Speech Processing*. Springer, 2015, pp. 27–38.

[19] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *in IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 333–336, 2002.

[20] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," *in AAAI Conference on Artificial Intelligence*, pp. 1919–1925, 2017.

[21] F. Plante, M. G, and A. W.A, "A pitch extraction reference database," *in Eurospeech*, pp. 837–840, 1995.

[22] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching." *in European Conference on SpeechCommunications and Technology*, pp. 1003—1006, 1993.

[23] J. D. O'Connor, *Better English Pronunciation*. Cambridge University Press, 1980.

[24] A. Saha, C. Yarra, and P. K. Ghosh, "Low resource automatic intonation classification using gated recurrent unit (GRU) networks pre-trained with synthesized pitch patterns." in *Interspeech*, 2019, pp. 959–963.

[25] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[26] A. Camacho, "SWIPE: A sawtooth waveform inspired pitch estimator," *Ph.D. thesis, University of Florida*, 2007.

[27] S. Gonzalez and M. Brookes, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[28] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[29] P. Mertens, "Polytonia: a system for the automatic transcription of tonal aspects in speech corpora," *Journal of Speech Sciences*, vol. 4, no. 2, pp. 17–57, 2014.

[30] ——, "Prosogram user's guide," 2020.