A frame selective dynamic programming approach for noise robust pitch estimation

Chiranjeevi Yarra,[1, a)] Om D. Deshmukh,[2] and Prasanta Kumar Ghosh[1]

[1]*Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, 560012, India*

[2]*Xerox Research Center India, Bangalore, 560103, India*

The principles of the existing pitch estimation techniques are often different and complementary in nature. In this work, we combine the complementary characteristics of two existing methods, namely, sub-harmonic to harmonic ratio (SHR) and sawtooth-wave inspired pitch estimator (SWIPE), to improve pitch estimation. Using variants of SHR and SWIPE, the proposed method, named FSDP, classifies all the voiced frames into two classes – the first class consists of the frames where a confidence score maximization criterion is used for pitch estimation, while for the second class, a dynamic programming (DP) based approach is proposed. Experiments are performed on speech signals separately from KEELE, CSLU and PaulBaghsaw corpora under clean and additive white Gaussian noise at 20, 10, 5, and 0dB SNR conditions using four baseline schemes including SHR, SWIPE and two DP based techniques. The pitch estimation performance of FSDP, when averaged over all SNRs, is found to be better than those of the baseline schemes suggesting the benefit of applying smoothness constraint using DP in selected frames in the proposed FSDP scheme. The VuV classification error from FSDP is also found to be lower than that from all four baseline schemes in almost all SNR conditions on three corpora.

[a] chiranjeeviy@iisc.ac.in

I. INTRODUCTION

Accurate estimation of pitch is useful in various applications including gender classification[1], emotion recognition[2], automatic intonation identification[3], automatic music transcription[4], query by humming[5], speech disorders identification[6] and source-filter model based speech coding systems[7,8]. The reliability of these applications depends on the accuracy of the pitch estimation. Typically pitch is considered as the fundamental frequency of the quasi-periodic speech signal perceived by the human auditory system[9–13]. An accurate pitch estimation for speech signal is non-trivial because – a) speech is not perfectly periodic due to non-stationary variations in the frequency and the amplitude[8], b) speech can be noisy, for example, in the case where the distance between the microphone and speaker is large, c) the signal to noise ratio (SNR) can be low[14].

In the literature, several estimation techniques[15–17] have used the dynamic programming (DP) to impose temporal continuity in the estimated pitch contour[12,18–24]. Typically, DP based approaches divide the input signal into frames and identify multiple pitch candidates for every frame. Often, these candidates are associated with a measure of confidence[18], referred to as confidence-score. These scores are considered in DP cost function for selecting the best candidate in a frame, which is declared as the estimated pitch at that frame.

The manner in which the pitch candidates and their confidence-scores are computed varies across different DP approaches. For example, in neutral network based approaches, probabilistic outputs of pitch candidate states are produced where a typical number of pitch candidates is approximately 68[22,23]. Among these, a deep neural network (DNN) based approach is shown to be effective in noisy speech, although it requires a large amount of data for training. In contrast to these data driven approaches, several knowledge based approaches are proposed with fewer number of pitch candidates. The robust algorithm for pitch tracking (RAPT) uses normalized cross correlation function (NCCF) to estimate multiple pitch candidates and their confidence-scores[20]. The yet another algorithm for pitch tracking (YAAPT) also uses NCCF to estimate multiple pitch candidates, but these candidates are further refined using spectral information[25]. The algorithm proposed by Ba et al., named BaNa, computes multiple pitch candidates by combining the approaches of harmonic ratio and cepstrum analysis[26]. The algorithm proposed by Gonzalez et al., named PEFA, uses a convoluted normalized periodogram to estimate multiple candidates and then pitch is estimated using DP[12]. While RAPT and YAAPT have been shown to perform well for clean and telephone channel speech respectively, PEFAC has been shown to perform better in low SNR conditions. However at higher SNRs and clean conditions, PEFAC does not have a satisfactory performance. In this work we propose a frame selective DP (FSDP) approach which works better with few number of pitch candidates in clean as well as in noisy conditions in both high and low SNR conditions. The proposed FSDP exploits speech characteristics in order to estimate pitch using small amount of training data.

Similar to the pitch candidates, the computation of the confidence-scores plays an important role in pitch estimation performance for both clean and noisy conditions. For example, RAPT has a robust confidence-score computation associated with each candidate, which could be the reason for it to have a better accuracy in clean case compared to other DP based algorithms. However RAPT involves the selection of many parameters on a training corpus, causing performance degradation across corpora as well as in noisy conditions[12]. To improve the performance under noisy conditions, PEFAC introduces another confidence-score computation that uses several parameters heuristically designed during training under noisy conditions.

Another critical factor for the performance of a DP based method is the weight given to the continuity constraint. While the continuity constraint often helps in correcting pitch halving and doubling errors, a large weight on the continuity constraint might introduce errors[12,27] by not recognizing gradual pitch transitions. Conversely, a weak continuity constraint may produce undesired fluctuations in the estimated pitch contour. These variations in the pitch estimation using RAPT and PEFAC are illustrated in Figure 1. In box 4, the estimated pitch from PEFAC is smoother than the ground truth, which has pitch transitions. This could be due to the strong continuity constraint in the DP. However, in boxes 1 and 2 (RAPT), the inaccurate transitions could be due to relaxed continuity constraint. In box 3, the pitch estimation error occurs due to the absence of the pitch candidate. This could be caused by inaccurate estimation or insufficient pitch candidates. Increasing the number of pitch candidates would require a carefully designed cost function for the DP to result in an accurate pitch contour. Hence, the effectiveness of a DP based approach depends on the degree of the continuity constraint and the accuracy of pitch candidates and their confidence-scores.

In this work, we propose a technique for computing pitch candidates and their confidence-scores by combining complementary characteristics of two existing methods, namely, sub-harmonic to harmonic ratio (SHR)[10] and sawtooth wave inspired pitch estimator (SWIPE)[9].

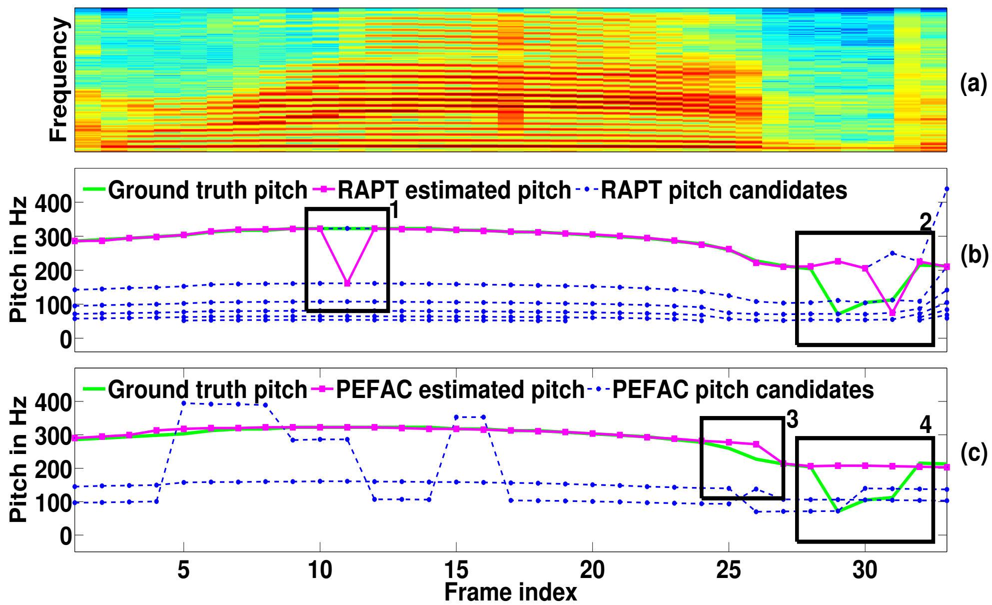

FIG. 1. An illustrative example describing two DP based techniques, namely RAPT and PEFAC – a) spectrogram of an exemplary voiced segment using an FFT of 1024 with frame shift and length of 10ms and 20ms respectively, b) pitch estimation by the RAPT algorithm, c) by the PEFAC algorithm. The erroneous regions in the estimated pitch contour are indicated with black boxes with box number at the top right corner.

Using these candidates, we employ a DP scheme to provide continuity in the pitch contour only in a few selected frames, called DP frames, unlike a typical DP method that works for all frames within a voiced segment. In the remaining non-DP frames, pitch is estimated using a maximal confidence-score criterion. We observe that SHR achieves a significant accuracy in the pitch estimation because it uses a good strategy for estimating reliable pitch candidates. However, it only computes two candidates, causing estimation error in cases where the ground truth pitch does not correspond to any of the candidates. We propose an extended candidate estimation strategy based on SHR to increase the number of pitch candidates, such that one of those candidates becomes more likely to correspond to the ground truth pitch. Similarly, we extend the confidence-score computation strategy in SWIPE by exploiting the window dependent properties (hanning window dependent kernel) and speech perception and production based properties. The latter includes equivalent rectangular bandwidth (ERB) frequency scale and decaying spectral envelope ($1/f$) similar to the glottal pulse spectrum. These confidence-scores are also used to automatically determine the DP and non-DP frames.

In addition to the proposed FSDP method for pitch estimation, we perform voiced-unvoiced (VuV) classification in each frame using the pitch candidate confidence-scores. Experiments for both pitch estimation and VuV classification are performed using three corpora: KEELE[28], CSLU[29] and PaulBaghsaw (PB)[30] in clean as well as noisy conditions with additive white Gaussian noise in 20, 10, 5 and 0dB SNRs. Gross pitch estimation (GPE)-20 error, root mean squared error (RMSE) and voiced and unvoiced (VuV) classification error are used as the evaluation metrics. We consider RAPT, PEFAC, SHR, and SWIPE as the baseline schemes. For pitch estimation, the proposed FSDP is found to achieve lower GPE-20 and RMSE compared to those of four baseline schemes, when the performance is averaged across all SNR conditions. FSDP performs better than all four baseline schemes for all three corpora in clean and in all SNR conditions, except for PaulBaghsaw corpus at 0dB SNR. For VuV classification, the proposed FSDP performs better than all four baseline schemes for all three corpora in clean as well as all SNR conditions except at 20dB SNR on CSLU corpus, where RAPT has the least VuV error.

II. PROPOSED FSDP APPROACH

The proposed FSDP approach has five stages, shown in Figure 2 and these stages are described using an exemplary voiced segment shown in Figure 3. The first stage computes pitch candidates ($p_t^k, 1 \leq k \leq K$) at the t-th frame, where K is the total number of pitch candidates. In the second stage the confidence-score $C_t(k)$ associated to each candidate is computed. In the third stage, a VuV decision is taken at each frame based on the confidence-scores $C_t(k)$, and using a support vector machine (SVM) classifier, which was learnt in the training. This VuV decision is used in the fourth and fifth stages. We consider contiguous estimated voiced frames as one estimated voiced segment. Figure 3 shows the pitch candidates from the first stage for $K = 2$ in an estimated voiced segment. In the fourth stage, all frames in each estimated voiced segment are divided into two sets – DP frames and non-DP frames based on $C_t(k), 1 \leq k \leq K$. In Figure 3, the pitch candidates of the non-DP and DP frames are shown using red squares and blue diamonds respectively. The fifth stage estimates pitch (magenta line in Figure 3) for both types of frames separately. For the non-DP frames, pitch is estimated using the following maximization criteria:

$$k^{opt} = \arg\max_k C_t(k); \qquad \hat{p}_t = p_t^{k^{opt}} \tag{1}$$

For the remaining frames, a DP based solution is used which selects one of the K pitch candidates in each frame such that the resultant pitch trajectory is maximally smooth within the segment.

It should be noted that the estimated unvoiced frames are not processed in the fourth and fifth stages. However, we use the maximization criteria in (1) to obtain pitch in the

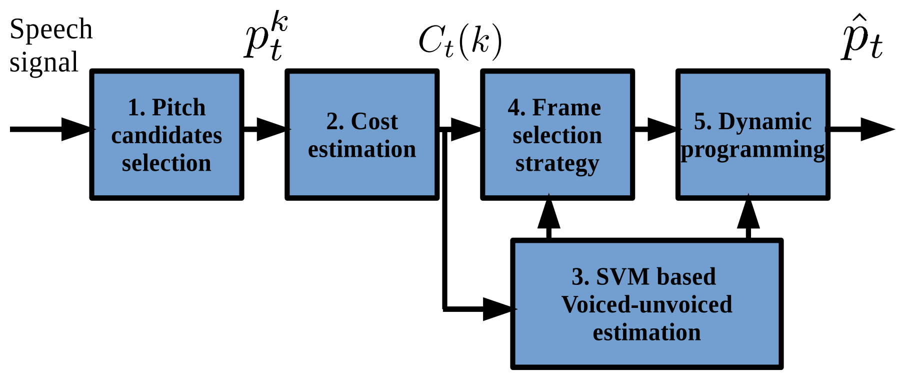

FIG. 2. Block diagram illustrating the steps of the FSDP method

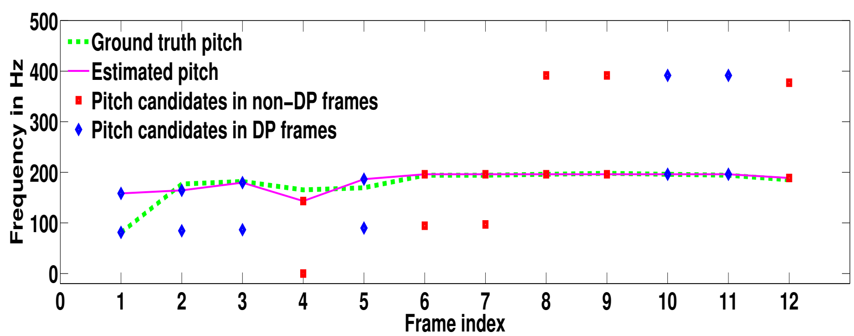

FIG. 3. An illustrative example explaining the proposed FSDP method

estimated unvoiced frames so that pitch is predicted in all frames of an utterance. This is done to obtain the pitch values at all ground truth voiced frames.

A. Pitch candidate selection

Pitch candidates are computed by following the two steps of the SHR method[10]. In the first step, we define $S_t(f)$ at the t-th frame as:

$$S_t\left(f\right) = \sum_{n=1}^{N} A_t\left(nf\right) - A_t\left(\left(n - \frac{1}{2}\right)f\right) \tag{2}$$

where $A_t(f)$ is the short time amplitude spectrum at the t-th frame and N is the maximum number of harmonics contained in $A_t(f)$[10]. The $S_t(f)$ measures the difference between amplitude sums at harmonic and at sub-harmonic components of the frequency f. This value is expected to be maximum at the pitch frequency because a typical spectrum of a periodic signal has high amplitudes at the harmonics of fundamental frequency and low amplitudes at the sub-harmonics. In the case of non-periodic signals, for example an unvoiced sound, the sum of the spectrum at the sub-harmonics would be relatively higher compared to that of a periodic signal and, hence, the $S_t(f)$ might not be as high as that of voiced (periodic) speech signal.

We observe that for some voiced speech segments the maximum of $S_t(f)$ may not correspond to the pitch frequency. Hence, pitch estimation based on a strategy that selects the frequency by maximizing $S_t(f)$ would introduce errors. We observe that most of these errors are pitch halving and doubling, which are also common source of errors in most of the existing pitch estimation methods[19]. This suggests that the candidate pitch frequency could be obtained by multiplying frequency corresponding to the highest peak of $S_t(f)$ with integer powers of 2.

In the second step, based on the above observation, we compute K pitch candidates as:

$$p_t^k = \begin{cases} \underset{f}{\arg\max}\ S_t(f) & \text{for } k = \lceil \frac{K}{2} \rceil \\ \underset{p_{t,Left}^{k-\lceil \frac{K}{2} \rceil} \le f \le p_{t,Right}^{k-\lceil \frac{K}{2} \rceil}}{\arg\max} S_t(f) & \text{for } k \neq \lceil \frac{K}{2} \rceil \end{cases} \tag{3}$$

where p_t^k is k-th pitch candidate at t-th frame for $k \in \{1, 2, ..., K\}$, $\lceil \frac{K}{2} \rceil$ is the smallest integer greater than $\frac{K}{2}$ and $p_{t,Left}^{k-\lceil \frac{K}{2} \rceil}$ and $p_{t,Right}^{k-\lceil \frac{K}{2} \rceil}$ are equal to $\left(1 - \frac{1}{16}\right) = 0.9375$ and

$\left(1+\frac{1}{16}\right) = 1.0625$ times $2^{k-\lceil \frac{K}{2} \rceil} p_t^{\lceil \frac{K}{2} \rceil}$ respectively. In particular, for $1 \leq k < \lceil \frac{K}{2} \rceil$, p_t^k includes the frequencies around the sub-harmonics (negative integer powers of 2) of $p_t^{\lceil \frac{K}{2} \rceil}$ that fall within the frequency band ranging from $p_{t,Left}^{k-\lceil \frac{K}{2} \rceil}$ to $p_{t,Right}^{k-\lceil \frac{K}{2} \rceil}$, based on $\frac{1}{6}$ octave band at each candidate, which is a linear approximation for the critical bands of the ear[31,32]. The $\frac{1}{6}$ octave band at $2^{k-\lceil \frac{K}{2} \rceil} p_t^{\lceil \frac{K}{2} \rceil}$ is equal to $(2^{\frac{1}{6}} - 1) \approx 0.125$ times $2^{k-\lceil \frac{K}{2} \rceil} p_t^{\lceil \frac{K}{2} \rceil}$, which is equal to $p_{t,Right}^{k-\lceil \frac{K}{2} \rceil} - p_{t,Left}^{k-\lceil \frac{K}{2} \rceil}$. Similarly, for $k > \lceil \frac{K}{2} \rceil$, p_t^k includes the frequencies around the harmonics (positive integer powers of 2) of $p_t^{\lceil \frac{K}{2} \rceil}$. We do not compute p_t^k beyond the typical pitch frequency range (50-550Hz). Hence the value of K is upper bounded by the total number of pitch candidates within the pitch range. The value of K is learnt in the training stage and is kept fixed for all the frames during the estimation of pitch and VuV decisions.

B. Candidate confidence-score computation

We modify the confidence-score computation steps in SWIPE[9] and define the confidence-score ($C_t(k)$) associated with each pitch candidate as:

$$C_t(k) = \frac{\sum_{f'} \Phi\left(p_t^k, f'\right) \sqrt{\Lambda_t\left(p_t^k, f'\right)} \frac{1}{\sqrt{f'}}}{\left|\Phi^+\left(p_t^k, f'\right) \frac{1}{\sqrt{f'}}\right| \left|\sqrt{\Lambda_t\left(p_t^k, f'\right)}\right|} \tag{4}$$

where, $\Lambda_t(p_t^k, f')$ is the amplitude spectrum of a windowed speech signal at the t-th frame with frequency index f'. The amplitude spectrum ($\Lambda_t(p_t^k, f')$) is computed for every pitch candidate p_t^k using Hanning window of size equal to $\frac{8}{p_t^k}$. $\Phi^+(p_t^k, f')$ is the positive part of the kernel $\Phi(p_t^k, f')$[9], which is defined for every pitch candidate p_t^k with frequency index f'

as $\sum_{i \in \{1\} \cup P} \Phi_i\left(p_t^k, f'\right)$ where P is the set of prime numbers and $\Phi_i\left(p_t^k, f'\right)$ defined as:

$$\Phi_i\left(p_t^k, f'\right) = \begin{cases} \cos\left(2\pi \frac{f'}{p_t^k}\right), & \left|\frac{f'}{p_t^k} - i\right| < \frac{1}{4} \\ \frac{1}{2}\cos\left(2\pi \frac{f'}{p_t^k}\right), & \frac{1}{4} < \left|\frac{f'}{p_t^k} - i\right| < \frac{3}{4} \\ 0, & otherwise \end{cases}$$

As demonstrated by Camacho et al[9], the $C_t(k)$ is typically high at the pitch frequency when the window and the kernel are chosen appropriately.

C. VuV decisions estimation

We consider an SVM based classifier for estimating VuV decisions as a binary classification task. We obtain VuV decisions from the K candidate confidence-scores $C_t(k)$ belonging to each frame as a feature vector. Along with these feature vectors, we use ground truth VuV decisions labeled from the ground truth pitch values to train the SVM. In the labeling procedure, we consider the frames corresponding to zero pitch values as unvoiced and the remaining frames as voiced for all three corpora.

D. Frame selection strategy using nearest neighborhood

We observe that due to the mismatch between window and kernel choices, $C_t(k)$ could be high at a pitch candidate different from the correct pitch frequency. Thus, determining pitch frequency by finding the frequency corresponding to the highest confidence-score (SWIPE strategy) may not work uniformly well in all frames. We propose a method to automatically determine the frames (referred to as non-DP frames) where taking the frequency corresponding to the highest confidence-score would accurately estimate the pitch frequency. In the remaining frames (referred to as DP frames), we use DP for estimating pitch. For DP, confidence scores are not used; rather, only pitch candidates are used. This helps in overcoming the errors in the pitch estimated by SWIPE strategy in DP frames. Towards this, in the training stage, we define two groups of pitch candidates – 1) the pitch candidate frequencies lying within $\pm 20\%$ of the ground truth pitch called required pitch candidates (RPCs); 2) other pitch candidates (non RPCs) away from (more than 20%) the ground truth pitch. We refer to the voiced frames corresponding to RPCs with the highest confidence-score as non-DP frames and the remaining voiced frames as DP frames and consider them as ground truth DP and non-DP frames. In order to determine the DP and non-DP frames in testing stage, we propose a frame selection strategy in the following section.

In the frame selection strategy, each frame of a voiced segment is categorized into either a DP frame or a non-DP frame. For this, we utilize the confidence-score associated with each candidate in developing the frame selection strategy. We use the confidence-scores of all pitch candidates as K-dimensional feature vector and pose the frame selection as a binary classification problem – non-DP frames as one class and DP frames as another class. The classification is done using the nearest neighborhood (NN) classifier[33,34] where r-nearest neighbors are computed based on the Euclidean distance. The parameter r is learnt during the training phase.

Algorithm 1 Pitch contour estimation algorithm based on DP

Initialization: $\mathscr{K} = \{1 : K\}$, $T =$ length of voiced segment

for each voiced segment **do**

  Initialization: $D_1(i) = 0 \;\; \forall \;\; i \in \mathscr{K}$

  for each frame t from 2 to T **do**

    $\forall \;\; i \in \mathscr{K}$

    if $t \in$ DP frames **then**

$$D_t(i) = \min_{j \in \mathscr{K}} \left\{ D_{t-1}(j) + \left(p_t^i - p_{t-1}^j \right)^2 \right\}$$

$$k_t(i) = \arg\min_{j \in \mathscr{K}} \left\{ D_{t-1}(j) + \left(p_t^i - p_{t-1}^j \right)^2 \right\}$$

    else

$$k^{opt} = \arg\max_{j \in \mathscr{K}} \{ C_t(j) \}$$

$$p_t^i = p_t^{k^{opt}}$$

$$D_t(i) = \min_{j \in \mathscr{K}} \left\{ D_{t-1}(j) + \left(p_t^i - p_{t-1}^j \right)^2 \right\}$$

$$k_t(i) = \arg\min_{j \in \mathscr{K}} \left\{ D_{t-1}(j) + \left(p_t^i - p_{t-1}^j \right)^2 \right\}$$

    end if

  end for

  Back tracking: $\eta_T = \arg\min_{i \in \mathscr{K}} \{ D_T(i) \}, \hat{p}_T = p_T^{\eta_T}$

  for each frame t from $T - 1$ to 1 **do**

$$\eta_t = k_{t+1}(\eta_{t+1})$$

$$\hat{p}_t = p_t^{\eta_t}$$

  end for

end for

E. Dynamic programming

Most of the pitch estimation algorithms are prone to octave errors, in which the estimated pitch contour has abrupt transitions and differs from the original pitch by a factor of two or a half[12,19]. However, realistic pitch contour does not vary such abruptly and pitch variation across frames is, in general, smooth in nature[12,19,27]. In order to avoid these abrupt jumps due to erroneous pitch estimates, we incorporate a temporal continuity constraint to estimate the pitch in the DP frames. The continuity constraint is implemented using DP approach[35,36] with the Euclidean distance as an objective measure. The objective function involved in the DP approach is given by

$$\hat{p}_t = \underset{p_t; t \in \mathcal{F}}{\arg\min} \sum_t (p_t - p_{t-1})^2 \tag{5}$$

$$\text{such that } \hat{p}_t = p_t^{k^{opt}} \; \forall \; t \in \text{non-DP frames}$$

where $\mathcal{F}$ is a set of frames in a voiced region. The detailed algorithmic steps for solving (5) are provided in Algorithm 1.

III. DATABASE

We use KEELE[28], CSLU[29], and PaulBaghsaw (PB)[30] corpora for all experiments in this work. Table I shows the details of the three corpora and the number of recordings considered in our experiments. In the experiments, we consider only the sentences belonging to both the male and the female subjects from all three corpora and exclude the sentences belonging to the children. In all three corpora, each spoken utterance has been recorded simultaneously with a laryngograph signal, which is used to compute the reference pitch considered as the ground truth. KEELE database consists of utterances from five male, five female and five children speakers reading "The North wind story". CSLU database consists of 50 phonetically rich sentences spoken by seven male and five female speakers. These sentences have been collected from the TIMIT and Harvard Psychoacoustic corpora[37]. Each speaker has uttered every sentence in three different contexts. PB database consists of 50 sentences spoken by one male and one female speakers.

TABLE I. Details of the three corpora used in the experiments in this work

		KEELE	CSLU	PB
Number of sentences	Overall	15	1800	100
	considered	10	1800	100
Number of speakers	male	5	7	1
	female	5	5	1
	children	5	–	–
Availability of laryngograph		Yes	Yes	Yes

IV. EXPERIMENTAL RESULTS

A. Experimental setup

We compare the performance of the proposed FSDP for pitch estimation and VuV classification with four existing methods (SHR, SWIPE, RAPT and PEFAC) using speech signal in clean condition and in additive white Gaussian noise in four SNR conditions: 20, 10, 5 and 0dB. KEELE, CSLU and PausBaghsaw (PB) corpora have been used for this purpose. Among the four existing methods, Matlab implementations of the four methods, namely, SHR, SWIPE, RAPT and PEFAC are directly available[38–40] and are used for the comparison. The gross pitch error (GPE-20) and root mean squared error (RMSE)[14] are used as the metrics for comparing the performance of pitch estimation using different methods. The GPE-20 is computed as $100 \times \frac{N_{err}}{N_v}$, where, N_{err} is the total number of erroneous 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. Ground truth pitch is computed from the laryngograph signal available with individual corpus. Both GPE-20 and RMSE are computed by discarding the estimated pitch at the boundary frames (first and last frame) in every voiced segment. The parameters K and r are learnt using ground truth VuVs separately for each corpus from randomly chosen 20% data in clean condition among which 75% of the data is used for the training and the remaining used for the development. Among K and r, first, we obtain the best K which results in the least GPE-20 error on the entire 20% data. Then, the best r is learnt using the best K considering the errors computed on the development set. The parameters corresponding to the least GPE-20 error on a corpus are used to estimate pitch within the corpus and across several other corpora in clean and noisy conditions using estimated VuVs to examine the generalizability of the proposed method.

The performance of VuV classification using different methods are compared using the classification error[41]. We use SVM classifier with RBF kernel for the classification task with the complexity parameter (C) equal to 1.0 and with kernel coefficient (γ) equal to 1/number of features. SVM classifier is implemented using Scikit-learn[42]. We train the SVM using a training set identical to that for learning the parameters in the pitch estimation task. These trained SVM models are used to estimate VuV decisions within the corpus and across the corpora in clean and noisy conditions. During comparison, we use readily available VuV decisions from all four existing methods except SWIPE for which classification error is not reported.

The performance of the proposed FSDP method depends on the accuracies in the estimation of DP & non-DP frames and VuV decisions. To understand the effect of each of these factors on the overall performance, we present the results in three sub-sections in Section IV B. Section IV B 1 discusses the pitch estimation accuracy with ground truth DP & non-DP frames and VuV decisions. Section IV B 2 discusses the effect of estimated DP & non-DP frames on the overall performance. Similarly, Section IV B 3 explains the effect of estimated DP & non-DP frames and VuV decisions. Following this, we analyze the reasons for a better performance using the proposed FSDP methods over four baseline schemes in two sub-sections – IV B 4 and IV B 5. For this analysis in Section IV B 4 and IV B 5, the benefit of FSDP are highlighted by comparing with SHR & SWIPE and with RAPT & PEFAC respectively. Note that, the performance of the proposed method also depends on the accuracy of the pitch candidates and their confident scores, which is discussed in Section IV B 6. Finally, in Section IV B 7, we present the accuracy of VuV classification.

TABLE II. GPE-20 obtained using the FSDP with ground truth DP and non-DP frames. A bold entry for a corpus and noise condition indicates the least GPE-20 among different K.

		$K = 2$	$K = 3$	$K = 4$
Clean	KEELE	0.79	**0.74**	0.75
	CSLU	1.16	**0.91**	0.95
	PB	1.34	1.33	**1.25**
20dB	KEELE	0.85	**0.81**	0.81
	CSLU	1.18	**0.92**	0.96
	PB	1.36	1.34	**1.26**
10dB	KEELE	1.20	**1.18**	1.20
	CSLU	1.49	**1.14**	1.20
	PB	1.90	1.89	**1.78**
5dB	KEELE	2.09	1.99	**1.97**
	CSLU	2.11	**1.67**	1.73
	PB	3.15	3.13	**2.92**
0dB	KEELE	5.29	5.08	**4.86**
	CSLU	3.88	3.16	**3.14**
	PB	6.25	6.21	**5.56**

TABLE III. GPE-20 obtained using the FSDP within and across all the three corpora using corpus specific parameters K and r learnt on the development set.

		FSDP		
		KEELE	CSLU	PB
clean	KEELE	**0.79**	1.04	1.17
	CSLU	1.61	**1.52**	1.74
	PB	1.49	1.45	**1.36**
20dB	KEELE	**1.12**	1.15	1.24
	CSLU	1.65	**1.57**	1.79
	PB	1.45	1.49	**1.36**
10dB	KEELE	**1.56**	1.60	1.67
	CSLU	**2.01**	2.02	2.15
	PB	2.03	2.02	**1.92**
5dB	KEELE	2.91	**2.73**	2.93
	CSLU	**2.73**	2.75	2.87
	PB	3.61	3.37	**3.21**
0dB	KEELE	6.90	**6.48**	6.59
	CSLU	4.78	**4.73**	4.88
	PB	7.57	6.88	**6.42**

B. Results and discussions

1. GPE-20 using FSDP with ground truth DP & non-DP frames and VuV decisions

Frame selection is one of the key components in the proposed FSDP approach. An error in frame selection causes errors in pitch values estimated using FSDP. Hence, we first compute the GPE-20 using FSDP where we use the ground truth DP and non-DP labels and VuV decisions (i.e., no errors due to either automatic frame selection or VuV classification). This could be used as the lower bound on the GPE-20 of the FSDP scheme. Table II shows these GPE-20 values computed on entire data from three corpora under clean and all noisy conditions for $K \in \{n; 2 \leq n \leq 4\}$. It is clear from Table II that the least GPE-20 increases with decreasing SNR. It also varies across different corpora. From the table, it is observed that the best K (corresponding to the least GPE-20) is 3 in clean, 20dB and 10dB SNR conditions and 4 in 0dB SNR conditions for KEELE and CSLU corpora. For PB, the best K is found to be 4 in clean and all noisy conditions. This indicates that the best K varies even within a corpus in clean and all noisy conditions; it also varies across three corpora. However, we consider the best K obtained in clean condition for each corpus to find the best choice of parameter r for NN based frame selection strategy.

2. GPE-20 using FSDP with estimated DP & non-DP frames and ground truth VuV decisions

The best choice of r is obtained for the frame selection strategy separately for each corpus using ground truth VuV decisions. We find the best choice of the parameter $r \in \{1 + 2n; 0 \leq n \leq 12\}$) on the development set for clean condition using GPE-20. The parameters K and r corresponding to the minimum GPE-20 are found to be (3 and 1), (3 and 21) and (4 and 1) for KEELE, CSLU and PB corpora respectively. From the optimal choice of r, it is observed that the parameter value changes in a corpus dependent manner.

Table III shows GPE-20 values on the entire data from all three corpora separately in clean and noisy conditions using K and r learnt for each corpus. In the table, each column indicates the corpus that is used for optimizing the parameters. It should be noted that the parameters are optimized for clean conditions. The diagonal entries (shaded regions in every $3{\times}3$ sub tables in Table III) indicate the GPE-20 values within the corpus (matched development and test corpora) and the off-diagonal values indicate errors across corpora (mis-matched development and test corpora). Bold entry for each corpus (every row) in Table III indicates the least GPE-20 value among all columns, which indicates the best development set. From the table, it is interesting to observe that the least errors are not confined to diagonal entries only, particularly at low SNR.

TABLE IV. Comparison of pitch estimation and VuV classification performance of RAPT, PEFAC, SHR, SWIPE and FSDP. The performance of different methods is compared using GPE (%) and RMSE (%).

		KEELE			CSLU			PB		
		GPE-20	RMSE	VuV	GPE-20	RMSE	VuV	GPE-20	RMSE	VuV
Clean	RAPT	2.85	17.27	8.99	4.39	20.85	6.77	2.24	34.93	10.40
	PEFAC	12.16	41.21	12.68	5.93	27.99	10.12	3.39	15.79	9.32
	SHR	1.73	13.16	12.72	2.63	17.12	13.51	1.93	11.14	8.13
	SWIPE	4.31	21.69	–	3.41	22.34	–	2.52	15.95	–
	FSDP	**0.89**	**8.90**	**6.43**	**1.65**	**13.31**	**5.91**	**1.50**	**9.01**	**7.29**
20dB	RAPT	4.07	21.18	6.30	5.12	24.99	**4.43**	3.94	16.12	7.49
	PEFAC	12.35	46.58	12.81	6.09	28.26	10.31	3.39	16.17	9.25
	SHR	1.87	12.95	8.22	2.73	17.54	5.82	1.99	11.56	6.54
	SWIPE	4.68	21.94	–	3.78	22.63	–	2.83	16.99	–
	FSDP	**1.17**	**9.91**	**6.12**	**1.69**	**13.54**	5.45	**1.48**	**8.85**	**6.51**
10dB	RAPT	16.48	35.51	9.46	14.81	34.37	6.31	16.19	27.31	7.75
	PEFAC	11.91	41.13	13.79	6.56	28.28	11.19	3.72	16.02	10.02
	SHR	2.50	14.97	15.01	3.42	19.31	11.07	2.46	12.08	9.05
	SWIPE	8.12	28.56	–	6.02	26.44	–	4.42	21.02	–
	FSDP	**1.59**	**11.39**	**7.09**	**2.05**	**14.56**	**5.58**	**2.00**	**9.60**	**5.87**
5dB	RAPT	31.70	53.35	17.03	24.07	46.78	10.94	25.39	42.36	10.47
	PEFAC	12.69	39.47	15.00	7.24	28.10	12.11	4.58	17.22	11.01
	SHR	4.24	18.28	26.48	4.59	21.50	21.42	3.87	14.45	17.17
	SWIPE	15.07	39.10	–	10.72	33.62	–	8.09	28.42	–
	FSDP	**2.79**	**13.43**	**9.73**	**2.79**	**16.01**	**6.82**	**3.39**	**12.12**	**6.69**
0dB	RAPT	59.77	75.76	30.62	48.67	69.38	23.21	51.63	69.01	20.07
	PEFAC	14.65	37.73	16.90	8.53	28.28	13.41	**6.38**	19.28	12.00
	SHR	8.26	23.99	41.10	7.52	26.23	36.24	7.90	20.38	28.27
	SWIPE	30.68	55.55	–	23.33	48.80	–	20.99	45.80	–
	FSDP	**6.56**	**18.81**	**16.26**	**4.77**	**19.15**	**10.61**	6.91	**17.18**	**10.87**

3. Comparison of GPE-20 and RMSE from FSDP and baseline schemes

Once the corpus and the SNR for a given test utterance is known, an accurate pitch contour could be achieved by using the parameters (K and r) corresponding to the least GPE-20 values (marked in bold) in Table III. However these corpus dependent parameters and the corresponding GPE-20 values might not be generalizable for unseen data. So, it may not be fair to compare these corpus and SNR specific GPE-20 values with the GPE-20 values computed using four baseline methods across all corpora and SNR conditions. Hence in FSDP, we consider one parameter set for frame selection strategy across all corpora and SNR conditions. This parameter set corresponds to the least GPE-20 value on the entire data among all corpora and SNR conditions (marked in blue in Table III). K and r in this parameter set are found to be the ones learnt on KEELE, i.e., K=3 and r=1. Using these parameters, we estimate VuV decisions using the SVM model learnt on KEELE with K=3. Following this, GPE-20 and RMSE are computed for all three corpora.

Table IV shows the GPE-20 and RMSE values obtained on the three corpora using the proposed FSDP and four baseline methods (RAPT, PEFAC, SHR and SWIPE) at various noisy levels and clean condition.In addition, we consider all frames as DP frames (i.e., no frame selection) in Equation 5 and compute the GPE-20 and RMSE to analyze the benefit of frame selection in FSDP scheme. However, pitch estimation using all frames as DP frames results in very poor performance; hence not reported in the table. The best performance for each metric is indicated in bold for each corpus and SNR condition. From the table, it observed that the proposed FSDP performs better than baseline methods for all three corpora in clean and all SNR conditions except at 0dB SNR in PB corpus (GPE-20 value), at which PEFAC performs better than FSDP. When averaged across clean and all noisy conditions, FSDP achieves the least average GPE-20 and RMSE errors (2.60 and 12.49, 2.59 and 15.31, 3.06 and 11.35) followed by SHR (3.72 and 16.67, 4.18 and 20.34, 3.63 and 13.92) for KEELE, CSLU and PB respectively. This implies that the strategies of SHR and SWIPE are complementary in nature and, when combined for computing pitch candidates and their confidence-scores as in FSDP, they achieve better pitch estimation accuracy compared to the individual ones in most of the cases. The improvement in the performance of FSDP over the four baseline methods is analyzed separately in two following subsections by comparing with – 1) SHR and SWIPE (the variants of which have been used in FSDP) 2) RAPT and PEFAC (DP based methods).

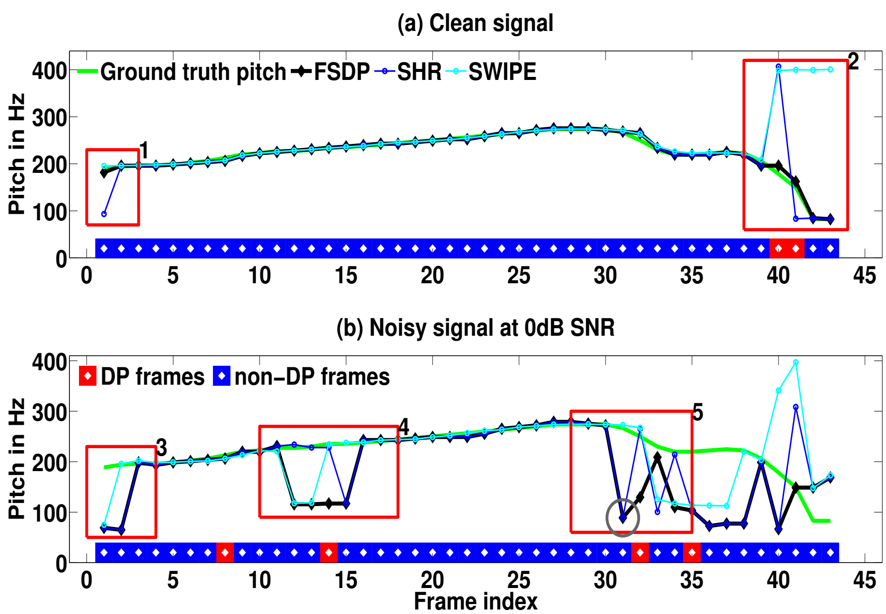

FIG. 4. Illustrative example describing the benefit of FSDP over SHR and SWIPE in a voiced segment. Red boxes 1, 2, 3, 4 and 5 are used to indicate the significant variations in the estimated pitch from the ground truth pitch. Red and blue horizontal patches indicate DP and non-DP frames respectively.

4. Comparison with SHR and SWIPE

Figure 4a and 4b show the estimated pitch trajectories for an exemplary voiced segment taken from the KEELE database in clean and noisy (SNR 0dB) conditions respectively. In

TABLE V. Comparison of erroneous frames (%) for both DP and non-DP categories obtained from FSDP in clean and all noisy conditions for all three corpora. All the percentages for each corpus are computed with respect to the total number of voiced frames

		estimated DP frames	EDPFs	ENDPFs	EDP_NDPFs (ENDPF′s)
KEELE	clean	0.89	0.25	0.65	0.02 (0.01)
	20dB	0.98	0.25	0.92	0.02 (0.01)
	10dB	1.26	0.32	1.27	0.04 (0.03)
	5dB	2.08	0.46	2.33	0.12 (0.10)
	0dB	2.64	0.74	5.82	0.34 (0.28)
CSLU	clean	0.70	0.17	1.48	0.03 (0.02)
	20dB	0.77	0.17	1.52	0.03 (0.02)
	10dB	1.05	0.22	1.84	0.05 (0.04)
	5dB	1.46	0.26	2.53	0.05 (0.04)
	0dB	2.48	0.44	4.33	0.18 (0.13)
PB	clean	1.15	0.06	1.43	0.00 (0.00)
	20dB	1.29	0.09	1.38	0.00 (0.00)
	10dB	1.79	0.24	1.76	0.00 (0.00)
	5dB	1.92	0.41	2.98	0.03 (0.02)
	0dB	2.93	0.70	6.21	0.26 (0.22)

box-1 all methods estimate pitch correctly except the SHR. This indicates that original pitch could be one of the pitch candidates in SHR, but the selection criteria used in SHR has led to wrong estimation of the pitch. In box-2, where the ground truth pitch has large variation, the proposed FSDP estimates pitch more accurately compared to all other methods. The SWIPE estimates wrongly at most of the points, which could be due to the large amount of errors in SWIPE when the actual pitch has wide variations. This could be because SWIPE considers many pitch candidates for estimating the pitch. SHR has better pitch estimates than those of SWIPE but worse than those of FSDP. When ground truth pitch has wide variation, we observe that the estimates of the pitch candidates and their confidence-scores become less reliable. This causes the SHR and SWIPE to result in octave errors. We also observe that such unreliable frames often get classified as DP frames using the nearest neighborhood strategy. Since the DP in the proposed scheme does not directly use the confidence-scores of the pitch candidates in DP frames and rather uses estimated pitch from neighboring non-DP frames to compensate the octave errors, the accuracy in the estimated pitch in these unreliable frames improves by using FSDP.

From Figure 4b, it is observed that the estimation errors are more in 0dB SNR compared to the clean condition for all the methods. This observation is consistent with the overall performance degradation in Table IV from clean to 0dB SNR condition. The performance degradation of FSDP could be due to two reasons. The first reason is that the estimated DP frames are more (2 in Figure 4a and 4 in Figure 4b, as highlighted using red horizontal patches) in case of 0dB SNR than in the clean condition. Higher number of DP frames could cause a smooth pitch trajectory even in frames with large ground truth pitch variations, and thereby resulting in a lower performance at 0dB SNR. The percentage of such DP frames that cause errors in the pitch estimation, called erroneous DP frames (EDPFs), are listed in the fourth column of Table V across all three corpora in clean and all noisy conditions. From the table, it is observed that the DP frames and EDPFs increase from clean to 0dB SNR condition for all three corpora. This implies that more DP frames result in more EDPFs, and, hence, the performance could degrade from clean to 0dB SNR.

The second reason for poor performance of FSDP in low SNR condition could be a large number of non-DP frames which result in pitch estimation errors at 0dB SNR, called erroneous non-DP frames (ENDPFs), (0 in Figure 4a and 15 in Figure 4b). In the entire set of ENDPFs, a subset of ENDPFs, indicated as ENDPF′ (13-th, 15-th and 31-st frames in Figure 4b), introduces pitch estimation errors in the neighboring DP frames due to the smoothing constraint in DP. For illustration, consider the 31-st frame marked in gray circle in Figure 4b in the box-5. This frame is classified as a non-DP frame (but it is ENDPF) by the nearest neighborhood frame selection strategy. Because of this, FSDP estimates the pitch values incorrectly at the neighboring DP frame (32-nd) in box-5 by following a wrong smooth trajectory. ENDPFs and ENDPF′ are listed in the fifth and sixth column (in brackets) of Table V. The percentage of such DP frames that results in pitch estimation errors due to ENDPF′s (indicated as EDP_NDPFs) are listed in the sixth column of Table V. From the table, it is observed that EDP_NDPFs are more than ENDPF′s for all three corpora. This indicates that the number of frames with estimated erroneous pitch is more for every ENDPF′ than that for every remaining ENDPFs. From the table, it is also observed that ENDPFs as well as ENDPF′s gradually increase from clean to 0dB SNR for all three corpora. Hence the additional pitch estimation errors by the ENDPFs along with pitch estimation errors by the EDPFs could result in further performance degradation. These observations from EDPFs and ENDPFs are consistent with the performance degradation of FSDP in Table IV for all three corpora.

5. Comparison with RAPT and PEFAC

From Table IV, it is observed that the GPE-20 of RAPT varies largely from clean to 0dB SNR condition compared to all other methods for all three corpora. This observation is consistent with the experimental findings by Gonzalez and Brookes[12]. The worst performance of RAPT at 0dB SNR could be due to the increase in incorrect pitch candidates by

NCCF. PEFAC performs worse in the clean case but better in the noisy case compared to RAPT. This is because it was designed specifically for noisy signal with low SNRs. However, FSDP performs better in both clean and noisy conditions in almost all cases. This superior performance could be because FSDP performs DP only in the selected frames with few pitch candidates (optimal $K = 3$) using a few parameters (K and r).

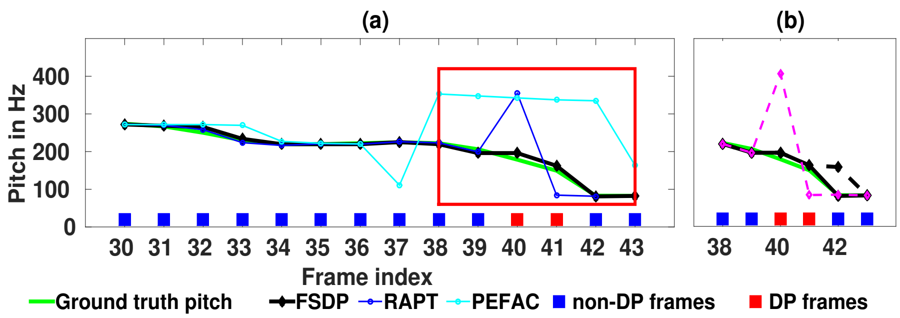

FIG. 5. Illustrative part of the voiced segment used in Figure 4 describing – a) the benefit of FSDP over RAPT and PEFAC b) the benefit of DP and non-DP frames in FSDP. The segment within red rectangular box in Figure 5a is shown in Figure 5b. The dotted black and magenta lines in Figure 5b indicate the estimated pitch trajectories when 40-th, 41-st and 42-nd frames are all DP frames and non-DP frames respectively.

Figure 5a shows the pitch trajectories obtained using RAPT, PEFAC and FSDP in clean condition for an exemplary voiced segment used in Figure 4. From Figure 5a, it is observed that PEFAC estimates incorrect pitch in the region, indicated by the red box in the figure, could be due to wrong pitch candidates in the highlighted region that result in a smooth trajectory away from the ground truth pitch. In the same region, RAPT estimates wrong pitch due to large deviations away from the ground truth pitch. However these errors are compensated in FSDP by using two different strategies in DP and non-DP frames – DP frames (40-th and 41-st) minimize the transitions and non-DP frames (38-th, 39-th, 42-nd, 43-th) allow pitch transitions without any smoothness constraint. Thus, the proposed FSDP allows pitch transitions as well as pitch smoothness in the right proportion using frame selection strategy thereby achieving better pitch estimation accuracy.

We elaborate these benefits with the help of Figure 5b, where, in addition to the pitch trajectory using FSDP, two other hypothetical trajectories (dotted-black and magenta) are shown when 40-th, 41-st and 42-nd frames are all assumed to be DP frames and non-DP frames respectively. It is clear that both trajectories suffer from pitch error either due to smoothness constraint (in 42-nd frame when all are assumed to be DP frames) or due to confidence-score maximization criterion (in 40-th and 41-st frames when all are assumed to be non-DP frames). However, providing smoothness constraint only in selected DP frames (as done in FSDP) results in an accurate pitch trajectory.

6. FSDP error analysis

Overall, pitch estimation errors using FSDP depend on the strategies used in DP and non-DP frames as well as the accuracy of the pitch candidates and their confidence-scores. We categorize these errors into three types. – 1) Absence of RPCs as pitch candidate selection strategy fails to detect them, 2) Estimated confidence-score associated with non-RPCs is the highest among all candidates (even when RPCs are present) due to the errors in candidate confidence-score estimation in the non-DP frames, 3) Selecting non-RPCs as the estimated

TABLE VI. Comparison of the number of GPE-20 frames belonging to three different types of errors occurred with different pitch candidates in clean and all noisy conditions for all three corpora.

		$K=2$			$K=3$			$K=4$		
		Absence of RPCs	with RPCs in non-DP	with RPCs in DP	Absence of RPCs	with RPCs in non-DP	with RPCs in DP	Absence of RPCs	with RPCs in non-DP	with RPCs in DP
KEELE	clean	0.41	0.23	0.32	0.26	0.39	0.24	0.26	0.39	0.32
	20dB	0.44	0.63	0.11	0.30	0.72	0.15	0.30	0.75	0.19
	10dB	0.81	0.63	0.18	0.67	0.71	0.21	0.67	0.81	0.21
	5dB	1.69	0.75	0.35	1.48	0.89	0.41	1.40	1.16	0.35
	0dB	4.73	1.23	0.55	4.35	1.51	0.71	4.06	1.84	0.68
CSLU	clean	0.87	0.75	0.04	0.57	0.98	0.09	0.57	1.12	0.05
	20dB	0.87	0.78	0.05	0.57	1.03	0.08	0.57	1.17	0.05
	10dB	1.14	0.89	0.06	0.78	1.18	0.09	0.77	1.32	0.06
	5dB	1.73	1.03	0.07	1.29	1.36	0.12	1.27	1.51	0.09
	0dB	3.40	1.36	0.12	2.75	1.81	0.19	2.69	2.03	0.17
PB	clean	1.25	0.20	0.00	1.20	0.26	0.00	1.11	0.25	0.01
	20dB	1.29	0.19	0.00	1.22	0.26	0.01	1.13	0.20	0.02
	10dB	1.83	0.20	0.00	1.77	0.25	0.00	1.64	0.25	0.04
	5dB	3.03	0.34	0.00	2.96	0.40	0.02	2.73	0.44	0.04
	0dB	5.97	0.89	0.03	5.85	1.01	0.04	5.26	1.09	0.11

pitch (even when RPCs are present) because of errors due to smoothing constraint using DP in DP frames.

Table VI shows the percentage of GPE-20 frames belonging to these three types of errors in FSDP for K =2,3,4 in clean and all noisy conditions for all three corpora. From the table it is observed that the errors due to the absence of RPCs are significant in most of the cases (especially at 5dB and 0dB SNR conditions for all three corpora). The errors due to the absence of RPCs are crucial in the proposed FSDP, since they determine the pitch estimation errors when there is no error in both frame selection strategy and pitch estimation strategy at DP and non-DP frames. We investigate the reason for this error in detail with the help of Figure 6 with K=2 using $S_t(f)$ for two exemplary voiced frames from the KEELE database. In the figure, the ground truth pitch frequency is indicated in green and the estimated pitch candidates are indicated in red. From Figure 6a, it is observed that the ground truth pitch frequency is closer to one of the pitch candidates. Hence, FSDP estimates the pitch accurately by choosing the correct pitch candidate. However in Figure 6b, both pitch candidates are far from the ground truth pitch which implies that the pitch candidate selection fails to estimate the RPCs. Hence FSDP fails to estimate the correct pitch. This underlines the importance of the pitch candidate selection method.

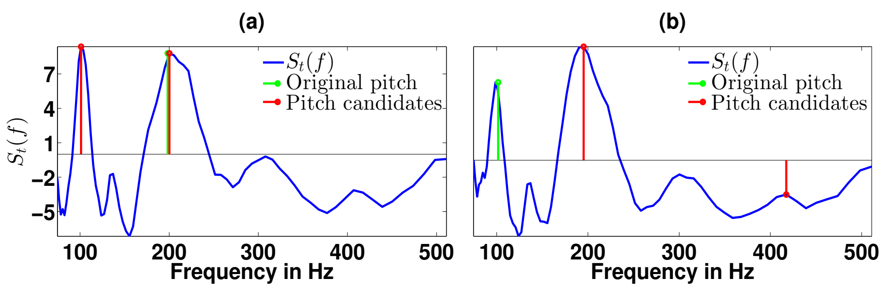

FIG. 6. Illustrative examples describing the importance of the pitch candidate selection method for $K = 2$ in which – a) the RPC is present b) the RPC is absent in the estimated pitch candidates

From Table VI, it is also observed that the errors due to the absence of RPCs reduce with increasing K for all SNRs and three corpora. This suggests that using more number of pitch candidates can reduce those errors. This could be because the search range of candidate selection method depends on K. Hence the RPCs which are missed with a low value of K can be detected with a high value of K. However, a high value of K does not guarantee a better pitch estimation accuracy due to increase in the second and third categories of error even when RPCs are present. This is also supported by the fact that the optimal K is found to be lower than 4 for all three corpora. Specifically, from Table VI, it is observed that the second type of errors consistently increases with K in clean and all SNR conditions for three corpora in almost all cases. This could be because the ambiguity in associating the maximum confidence-score with the RPCs increases with increasing in K. Hence, the pitch estimation method at non-DP frames fails to estimate the RPCs as a pitch.

The second and third categories of errors depend on the accuracy of the frame selection strategy and pitch estimation at DP and non-DP frames. We investigate these errors in the proposed FSDP considering three different choices of DP and non-DP frames – 1) ground truth DP and non-DP frames, 2) estimated DP and non-DP frames using the nearest neighborhood 3) all frames as non-DP frames (to highlight the importance of DP frames). We find the sum of second and third types of errors as 0.40, 1.01 and 1.04 respectively for the above mentioned three choices, when averaged across all five SNRs and all three corpora. Note that, considering all frames as DP frames results poor performance, hence not reported. For computing these errors, the parameters of the proposed FSDP are kept identical to those used in Table IV for all three corpora. It is observed that the average errors increase (monotonically) from the first choice to the third choice. The non-zero error in the first choice indicates that the errors are only due to incorrect pitch estimation at some of DP and non-DP frames. A higher error in the second choice compared to the fist choice indicates combined effect of the errors caused by the frame classification strategy and pitch estimation methods at DP and non-DP frames. Similarly, the highest error in the third choice indicates that the pitch estimation errors due to the errors in the frame selection strategy are less than those due to the pitch estimation strategy.

7. *VuV classification errors*

Table IV shows the VuV classification errors computed using FSDP and four baseline methods for all three corpora. In the table, a bold entry for a given corpus and SNR combination indicates the lowest VuV classification error. From the table, it is observed that the proposed FSDP has the least VuV error in clean and all noisy conditions on all three corpora except at 20dB SNR on CSLU corpus, where RAPT has the least VuV error. This indicates that the proposed FSDP method performs better than the four baseline schemes both in the pitch estimation and VuV classification tasks. It is interesting to notice that no single baseline scheme has consistently performed the best among all the baseline schemes across all noisy conditions on three corpora.

V. CONCLUSIONS

Realistic pitch trajectories are typically smooth in nature, but sometime they show large variation in pitch values in a short span of time. In this work, we propose FSDP approach for pitch estimation, which allows the estimated pitch trajectory to be smooth using DP only in a few selected frames (called DP frames) unlike a typical DP based method which forces the trajectory to be smooth over an entire voiced segment. In the remaining frames (called non-DP frames), FSDP approach allows large variation in the estimated trajectory by estimating pitch using a pitch candidate confidence-score maximization criterion where the candidates and their confidence-scores are computed using variants of SHR and SWIPE. These confidence-scores are used to automatically identify DP and non-DP frames. These confidence-scores are also used for VuV classification using an SVM classifier. Experiments with three corpora namely, KEELE, CSLU and PaulBaghsaw reveal that FSDP performs better than four baseline methods considered in this work for pitch estimation as well as VuV classification tasks.

The performance of the proposed FSDP method depends on the percentage of missing RPCs, reliability in estimating the pitch candidate confidence-scores, classification accuracy of DP and non-DP frames and effectiveness of smoothening constraint used in DP. Among all the errors, the percentage of missing RPCs is found to be crucial, since these errors determine the lower bound on the pitch estimation errors by the proposed method. Hence, further investigation is required to reduce the missing RPCs with an appropriate candidate selection strategy. In addition to this, the frame selection strategy needs to be improved. Most of the errors in the nearest neighborhood based frame selection strategy is due to the misclassification of DP and non-DP frames. Also, the computational cost involved in the frame selection strategy is quite high. This is because the nearest neighborhood classifier in the frame selection strategy computes a distance for each frame with all training samples. The training set for the DP and non-DP frame classification is also found to be imbalanced with a ratio of 1:100. Hence, a noise robust classifier with less computational complexity under large imbalanced training set would be effective.

VI. REFERENCES

REFERENCES

[1]M. Wang and M. Lin, "An analysis of pitch in chinese spontaneous speech," International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages (2004).

[2]I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," The Journal of the Acoustical Society of America **93**(2), 1097–1108 (1993).

[3]K. De Bot, "Visual feedback of intonation I: Effectiveness and induced practice behavior," Language and Speech **26**(4), 331–350 (1983).

[4]A. Askenfelt, "Automatic notation of played music: the VISA project," Fontes Artis Musicae 109–120 (1979).

[5]R. B. Dannenberg, W. P. Birmingham, G. Tzanetakis, C. Meek, N. Hu, and B. Pardo, "The musart testbed for query-by-humming evaluation," Computer Music Journal **28**(2), 34–48 (2004).

[6]E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," The journal of the Acoustical Society of America **71**(6), 1544–1550 (1982).

[7]G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, Vol. 2 (Walter de Gruyter, 1971).

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

[9] A. Camacho and J. G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” The Journal of the Acoustical Society of America **124**(3), 1638–1652 (2008).

[10] X. Sun, “Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio,” IEEE International Conference on Acoustics, Speech, and Signal Processing **1**, 333–336 (2002).

[11] A. De Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” The Journal of the Acoustical Society of America **111**(4), 1917–1930 (2002).

[12] 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 **22**(2), 518–530 (2014).

[13] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, “Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech,” IEEE Transactions on Speech and Audio Processing **13**(5), 776–786 (2005).

[14] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, “Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme,” IEEE Transactions on Audio, Speech, and Language Processing **20**(1), 322–335 (2012).

[15] W. Bauer and W. Blankenship, “Dyptrack–a noise-tolerant pitch tracker,” Dept. of Defence (NSA), Washington, USA, Unclassified Rep. NASL-S-210 (1974).

[16] H. Ney, “A dynamic programming technique for nonlinear smoothing,” IEEE International Conference on Acoustics, Speech, and Signal Processing **6**, 62–65 (1981).

[17]H. Ney, "Dynamic programming algorithm for optimal estimation of speech parameter contours," IEEE Transactions on Systems, Man and Cybernetics, **2**, 208–214 (1983).

[18]L. Sukhostat and Y. Imamverdiyev, "A comparative analysis of pitch detection methods under the influence of different noise conditions," Journal of Voice **29**(4), 410–417 (2015).

[19]M. Asgari and I. Shafran, "Improving the accuracy and the robustness of harmonic model for pitch estimation.," Proceedings Interspeech 1936–1940 (2013).

[20]D. Talkin, *A robust algorithm for pitch tracking (RAPT)* (W.B. Kleijin and K. K. Paliwal Eds. Amesterdam, The Netherlands: Elsevier, 1995), pp. 495–518.

[21]E. Azarov, M. Vashkevich, and A. Petrovsky, "Instantaneous pitch estimation based on RAPT framework," European Signal Processing Conference (EUSIPCO) 2787–2791 (2012).

[22]K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**(12), 2158–2168 (2014).

[23]H. Su, H. Zhang, X. Zhang, and G. Gao, "Convolutional neural network for robust pitch determination," IEEE International Conference on Acoustics, Speech, and Signal Processing 579–583 (2016).

[24]K. Han and D. Wang, "Neural networks for supervised pitch tracking in noise," IEEE International Conference on Acoustics, Speech, and Signal Processing 1488–1492 (2014).

[25]K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," IEEE International Conference on Acoustics, Speech, and Signal Processing **1**, 361–364 (2002).

[26] H. Ba, N. Yang, I. Demirkol, and W. Heinzelman, "BaNa: A hybrid approach for noise resilient pitch detection," IEEE Statistical Signal Processing Workshop (SSP) 369–372 (2012).

[27] L. Dolansky and P. Tjernlund, "On certain irregularities of voiced-speech waveforms," IEEE Transactions on Audio and Electroacoustics, **16**(1), 51–56 (1968).

[28] F. Plante, M. G, and A. W.A, "A pitch extraction reference database," Proceedings Eurospeech 95 837–840 (1995).

[29] A. Kain, "CSLU: Voices," Linguistic Data Consortium, Philedelphia, PA, USA (2006).

[30] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching.," Proceedings of the Third European Conference on SpeechCommunications and Technology 1003–1006 (1993).

[31] F. A. Everest, K. C. Pohlmann, and T. Books, *The master handbook of acoustics*, Vol. 4 (McGraw-Hill New York, 2001).

[32] H. Fletcher, *Speech and hearing in communication* . (D. van Nostrand, 1953).

[33] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," Information Theory, IEEE Transactions on **13**(1), 21–27 (1967).

[34] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," ACM Transactions on Mathematical Software (TOMS) **3**(3), 209–226 (1977).

[35] R. E. Bellman and S. E. Dreyfus, *Applied dynamic programming* (Rand Corporation, 1962).

[36]R. Bellman, *Dynamic programming* (Dover publications, 1957).

[37]A. B. Kain, "High resolution voice transformation," Oregon Health & Science University (2001).

[38]X. Sun, "Pitch determination algorithm," Software, available [Jan 2016] from http://in.mathworks.com/matlabcentral/fileexchange/1230-pitch-determination-algorithm/content/shrp.m .

[39]A. Camacho, "SWIPE pitch estimation algorithm," Software, available [Jan 2016] from http://www.cise.ufl.edu/ acamacho/publications/swipep.m .

[40]M. Brookes, "VOICEBOX: A speech processing toolbox for matlab. 2006," URL http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. Available online (2003).

[41]T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics.," Interspeech 1973–1976 (2011).

[42]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in python," Journal of Machine Learning Research **12**, 2825–2830 (2011).