Spectral refinement with adaptive window-size selection for voicing detection and fundamental frequency estimation

Nilesh Madhu
IDLab, Dept. Electronics & Information Systems
Ghent University - imec, Belgium
nilesh.madhu@ugent.be

Mohammed Krini
Signal Processing Laboratory
Aschaffenburg University of Applied Sciences, Germany
mohammed.krini@th-ab.de

***Abstract*—Spectral refinement (SR) offers a computationally inexpensive means of generating a refined (higher resolution) signal spectrum by linearly combining the *spectra* of shorter, contiguous signal segments. The benefit of this method has previously been demonstrated on the problem of fundamental frequency (F0) estimation in speech processing – specifically for the improved estimation of very low F0. One drawback of SR is, however, the poorer detection of voicing *onsets* due to the Heisenberg-Gabor limit on time and frequency resolution. This may also lead to degraded performance in noisy conditions. Transitioning between long- and short-time windows for the spectral analysis may offer a good trade-off in these situations. This contribution presents a method to *adaptively* switch between short- and long-time windows (and, correspondingly, between the short-term and the refined spectrum) for voicing detection and F0 estimation. The improvements in voicing detection and F0 estimation due to this adaptive switching is conclusively demonstrated on audio signals in clean and corrupted conditions.**

***Index Terms*—spectrum computation, fundamental frequency estimation, speech enhancement, DFT, spectral refinement**

I. Introduction

Audio and speech enhancement algorithms usually operate on sub-band representations of the signals [1], [2]. These are obtained by transforming overlapped, segmented (windowed) time-frames of the signals into the frequency domain using filterbanks. The parameters for the frequency (sub-band) transformation are generally selected to offer a good trade-off between frequency resolution and output latency. However, the windowing of the input frames before computing the sub-band representation introduces frequency overlap between the signals in adjacent sub-bands. Consequently closely separated harmonics occurring in voiced segments, cannot be easily separated. Thus, the compromise is usually detrimental to the performance of algorithms exploiting the fundamental frequency (F_0) of speech, especially when F_0 is low. Such algorithms include, among others, approaches that exploit the properties of speech for explicitly boosting the speech harmonics in voiced segments [3]–[5] and for restoring the phase consistency [6]. For such approaches, a reliable detection of voiced/unvoiced segments and an accurate estimation of the fundamental frequency F_0 is of paramount importance. Poor voiced/unvoiced detections and inaccurate F_0 estimates lead to annoying, audible artefacts in the enhanced speech, and degrades the overall performance of the enhancement schemes. Poor spectral resolution can also affect the performance of F_0 estimation approaches used, e.g., in [7], [8].

The spectral leakage and aliasing effects can be ameliorated by increasing the size of the data window. However, this leads to a larger overall delay in the signal path and as delay restrictions for hands-free telephony and in-car communications are strict, a longer data-window often means that these restrictions can not be fulfilled anymore [9], [10]. In most hands-free systems, therefore, the analysis and synthesis filterbanks and overlap-add or overlap-save based schemes that satisfy the different requirements (delay in signal path, aliasing properties, computation complexity, etc.) are tried and proven (legacy) implementations and can not be easily changed. Hence, the improved or higher-resolution spectrum must be computed on a separate data buffer for use in the analysis. A straightforward solution is, thus, to buffer the signal to the desired (longer) length and then compute the sub-band representation on this buffer. This requires, however, an additional implementation of the higher-order sub-band analysis and, consequently, additional computational cost.

In [11], the method of spectral refinement (SR) was first proposed, where the high-resolution spectrum corresponding to the use of a higher-order DFT was obtained by a *linear* combination of the current and preceding frames of the low-resolution short-term DFT spectra. This approach was shown to have the following advantages over the straightforward alternative: (a) it is computationally significantly less expensive; (b) it can be configured to produce either the full, high resolution spectrum or only compute the refined spectrum for a *subset* of desired frequencies – leading to a further reduction in computational expense, without compromising on the feature-extraction performance. This approach was generalised and further extended to the case of polyphase filterbanks in [12] and the benefit of this extension for the case of F_0 estimation was demonstrated in [13]. However, a shortcoming of using a limited-size data-window is the restriction imposed by the Heisenberg-Gabor limit (see e.g. [14]) namely, the impossibility of sharply localising a signal simultaneously in the temporal and spectral domain. For voicing detection and

F_0 estimation, when the voicing is *sustained* the temporal resolution is of less consequence and we benefit from the use of a longer data-window for more accurate F_0 estimation (better frequency resolution). However, for voicing *onsets*, since they are typically preceded by silence/unvoiced speech, using a long data-window generally gives a delay of several frames (depending on the overlap chosen) before voicing is detected. The use of a *short* data window is indicated in this case. Thus, the choice of a long or short data window should be made depending on whether we expect sustained voicing or not. In this paper we propose a method to switch *adaptively* between the short-term and the refined spectra for improved voicing onset detection and analyse the benefit of such adaptive switching compared to the case of always using a short term spectrum or the refined spectrum.

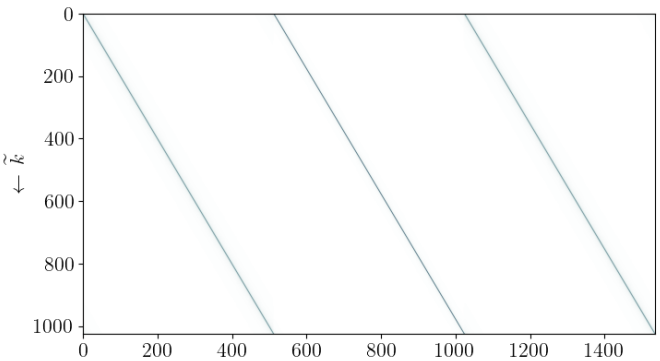
The rest of the paper is structured as follows: we first briefly describe the spectral refinement approach. Next, we indicate the problem with long data windows (i.e., using the refined spectra) and present our solution (adaptive switching). We then evaluate the proposed method on clean and corrupted speech and present our conclusions. Whereas we consider here the sub-band decomposition by the use of the discrete Fourier transform, the results may also be extended to the case of filterbanks (e.g. [12]).

II. Spectral refinement (SR)

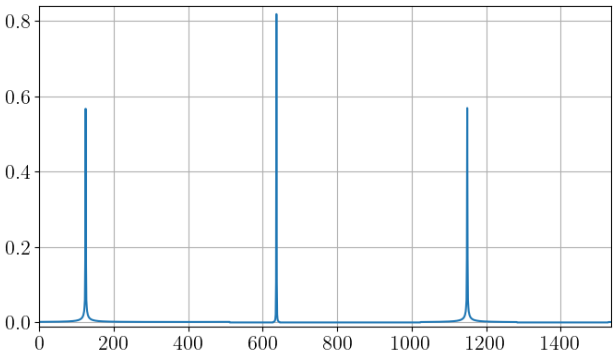
Let $y(\ell)$ be the discrete time-domain signal under consideration, obtained at a sampling frequency F_s. The short-term spectrum of this signal, obtained from an N-point DFT on windowed, overlapped signal segments, can be written as $\boldsymbol{Y}(n) = \big[Y(0,n), Y(1,n), \ldots, Y(N-1,n)\big]^T$. In this representation, $Y(k,n)$ corresponds to the complex spectral amplitude of the signal at frame n and the k-th discrete frequency $f_k = kF_s/N$. The idea behind spectral refinement is, then, to obtain the high-resolution spectrum $\widetilde{\boldsymbol{Y}}(n) = \big[\widetilde{Y}(0,n), ..., \widetilde{Y}(\widetilde{N}-1,n)\big]^T$, ($\widetilde{N} = IN$, $I \in \{2,3,...\}$) by a *linear combination* of successive N-point short-term spectra $\mathbf{Y}(n)$, $\mathbf{Y}(n-1)$,...,$\mathbf{Y}(n-L)$ of size N, where L is termed the *history* or *memory* for the refinement. This can be expressed mathematically as:

$$\widetilde{\mathbf{Y}}(n) = \mathbf{S}\left[\mathbf{Y}^T(n), \mathbf{Y}^T(n-1), \ldots, \mathbf{Y}^T(n-L+1)\right]^T, \quad (1)$$

where $\mathbf{S} \in \mathbb{C}^{\widetilde{N} \times LN}$ is the refinement matrix. It was shown in [11] that, by choosing the windows and frameshift (hop size) appropriately, a *sparse* matrix $\mathbf{S}$ is obtained, whereby computing $\widetilde{\mathbf{Y}}(n)$ by (1) is computationally *significantly less expensive* than computing it by a DFT on the $\widetilde{N}$-point time-domain signal. Figure 1 illustrates this sparse nature of the SR matrix. Further, the spectral refinement matrix can also be configured to produce the refined spectrum only for a *subset* of desired frequencies within the high-resolution spectrum – further reducing the computational cost. These properties make spectral refinement particularly interesting for speech processing systems with a legacy analysis/synthesis framework (and, thereby, with a fixed resolution) since it makes an arbitrary frequency resolution possible *without* introducing additional latency or significant computational overhead.



(a) Amplitude values of the SR matrix ($N = 512 \rightarrow \widetilde{N} = 1024$). The y-axis indicates the frequency supporting points in the higher resolution spectrum.



(b) Amplitude values of a particular row of the SR matrix above. Note that very few values are significant.

Fig. 1. Spectral refinement matrix for doubling the resolution from $N = 512$ to $\widetilde{N} = 1024$. The overlap between two adjacent frames in the short-time spectral analysis is 256 samples, leading to $L = 3$. This configuration yields an $\widetilde{N} \times LN = 1024 \times 1536$-dimensional $\mathbf{S}$, which is highly sparse.

III. Voicing detection and F_0 estimation with refined short-term spectra

There exists a broad variety of algorithms for estimating the fundamental frequency of a speech signal, such as the analysis in the cepstral domain [15], Bayesian approaches [8], or short-term auto-correlation based schemes [16].

For the purpose of illustration we consider, here, the latter approach as presented in [11], where the auto-correlation function is derived from the power-spectral density ($\Psi(k,n)$) of the signal. When computing the $\Psi(k,n)$ on the refined spectra $\widetilde{Y}(k,n)$, $\widetilde{\boldsymbol{Y}}(n)$ is first obtained (for the desired subset of frequencies) by applying SR as in (1) to the short-term spectra $\boldsymbol{Y}(n)$. The power spectral density (PSD) is subsequently approximated by the periodogram:

$$\widehat{\Psi}_{\widetilde{y}\widetilde{y}}(k,n) = |\widetilde{Y}(k,n)|^2 . \quad (2)$$

This estimate is subsequently divided by the spectral *envelope* $\overline{\Psi}_{\widetilde{y}\widetilde{y}}(k,n)$, which can be obtained, e.g., by the LPC [2], [17]:

$$\mathring{\Psi}_{\widetilde{y}\widetilde{y}}(k,n) = \frac{\widehat{\Psi}_{\widetilde{y}\widetilde{y}}(k,n)}{\overline{\Psi}_{\widetilde{y}\widetilde{y}}(k,n)} . \quad (3)$$

This operation results in the power spectral estimate of the so-called *excitation* signal which, in a voiced segment, consists

primarily of the spectral peaks at F_0 and its harmonics. Transforming $\mathring{\Psi}_{\tilde{y}\tilde{y}}(k,n)$ from (3) into the time-domain, we obtain an estimate of the short-term auto-correlation function (ACF) $\widehat{r}_{\tilde{y}\tilde{y}}(m,n)$, for different lags m. From this, the fundamental frequency $F_0(n)$ is determined by the position of the maximum within a selected range of lag-indices:

$$m_0(n) = \underset{m:m\in\{m_{\text{low}},m_{\text{high}}\}}{\text{argmax}} \left\{\widehat{r}_{\tilde{y}\tilde{y}}(m,n)\right\},$$
$$\text{and} \qquad F_0(n) = \frac{F_s}{m_0(n)}. \tag{4}$$

The limits m_{low} and m_{high} of the allowable range of lag-indices are determined by the typical range of F_0 for human voice (between 50 Hz and 300 Hz).

The auto-correlation function provides an additional feature to indicate the *reliability* of the voicing detection – the *normalised* ACF-value at $m_0(n)$:

$$\widetilde{\rho}(n) = \frac{\widehat{r}_{\tilde{y}\tilde{y}}(m_0(n),n)}{\widehat{r}_{\tilde{y}\tilde{y}}(0,n)}. \tag{5}$$

Then, if the signal segment is perfectly periodic (consistent, sustained voicing), we can expect $\widetilde{\rho}(n) \approx 1$. Thus, large values of $\widetilde{\rho}(n)$ in (5) may be interpreted as a high probability that the particular frame is voiced. With this interpretation, F_0 estimates are only considered for frames n where the $\widetilde{\rho}(n)$ exceeds a predefined threshold value, and such frames are demarcated as voiced frames. Similarly, the F_0 and the voicing feature $\rho(n)$ are computed for the short-term spectrum $Y(k,n)$.

IV. Limitation of Long Data-Windows

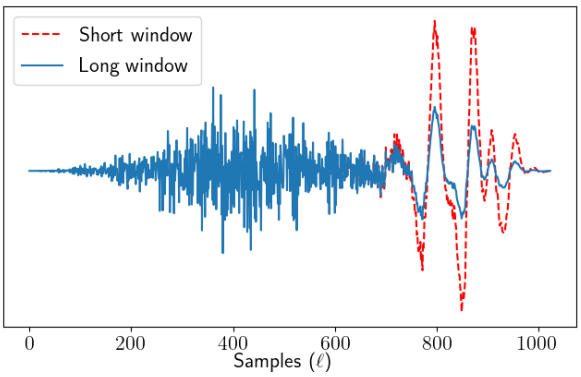
Figure 2(a) depicts a short data frame and the corresponding longer frame. The normalised auto-correlation for both these data windows is depicted in Figure 2(b). From Figure 2, the limitation of always using a long data window is evident when speech/voicing onsets occur. Another possible limitation occurs for noise corrupted speech where, in highly dynamic situations, always using a long data window for the F_0 and voicing estimates may not be advantageous. We suggest, therefore, a *switching* between the different window-lengths, picking the data-window most suitable at a particular frame.

V. Adaptive Switching of Windows

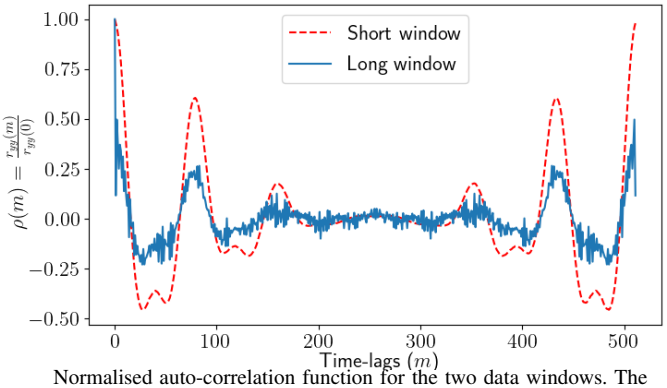
A straightforward approach to switch between the data windows (and, correspondingly, between the choice of the original and the refined spectrum) would be to select the data-window with the highest evidence of voicing. While this can be done by a straightforward comparison of $\rho(n)$ and $\widetilde{\rho}(n)$ as:

$$\text{Choose} \begin{cases} \widetilde{\mathbf{Y}}(n) \text{ if } \widetilde{\rho}(n) \geq \rho(n), \\ \mathbf{Y}(n) \text{ otherwise}, \end{cases} \tag{6}$$

this is not a good idea because the range of $\rho(n)$ and $\widetilde{\rho}(n)$ are quite different, with $\rho \leq \widetilde{\rho}$ in general, even in the presence of sustained voicing. This can be seen in the scatterplot in Figure 3, which plots ρ against $\widetilde{\rho}$ for a set of voiced segments extracted from clean speech. It may be seen from Figure 3 that while $\widetilde{\rho}$ gets close to 1 (likely in periods of sustained voicing), ρ is thresholded at a lower value for the same frames.



(a) A long and short data frame, at beginning of a voicing region.



(b) Normalised auto-correlation function for the two data windows. The voicing indication from the short window $\rho(n) \approx 0.70$ is reliable whereas that for the long window $\widetilde{\rho}(n) \approx 0.25$ is significantly less.

Fig. 2. A long and short data frame at the onset of voicing. The periodic component is well captured in the short frame, but a similar analysis of the longer data window would indicate a less-reliable detection of voicing. Here, even the F_0 estimate in the case of the long window would have a larger error compared to the short window (note the unambiguous peak in the ACF of the short window as compared to the noisier peak from the long window).

Thus, a direct comparison of the voicing feature as in (6) would tend to bias the decision in favour of long windows – which is undesirable. In fact, in terms of voicing indication, a 'lower' ρ of e.g. 0.75 would be a *stronger* indication of voicing compared to $\widetilde{\rho}$ of the same value. We therefore propose to first *individually transform* the voicing feature for each case into a measure that is *directly comparable*. This can be easily achieved by designing a *logistic regression classifier* (see e.g., [18]) separately for ρ and $\widetilde{\rho}$.

A. Logistic regression on ρ and $\widetilde{\rho}$ for probability of voicing

The logistic (or logit) model can be used to model the probability of a certain class in a *binary* classification problem, given the discriminatory input feature. For a general binary classification problem (i.e., 'Class 0' vs 'Class 1') based on an input feature vector $\mathbf{x}$, the output of a logistic regression classifier would be given as:

$$\mathcal{L}(\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} \in [0,1], \tag{7}$$

where $\mathbf{w}$ contains the *weights* of the trained model. The convention is to interpret the output $\mathcal{L}(\mathbf{x})$ as the *probability*

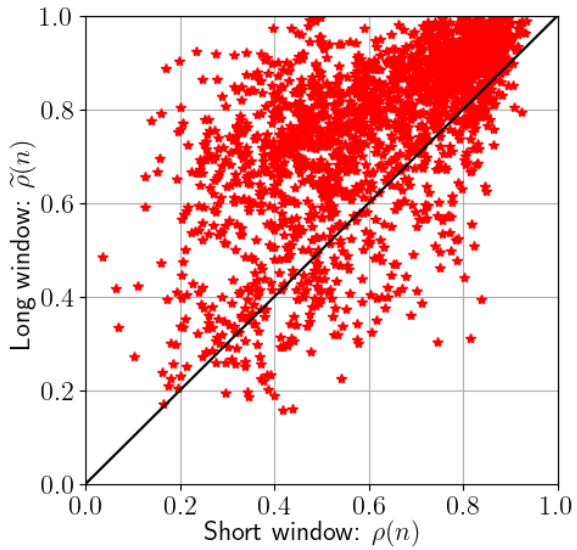


Fig. 3. Comparison of the voicing features when using the normal ($\mathbf{Y}$) or the refined ($\widetilde{\mathbf{Y}}$) spectra for detecting voicing and F_0. The voicing feature $\widetilde{\rho}$ corresponding to the refined spectrum (i.e., with the long window) is almost always higher than ρ, which corresponds to using the original short-term spectrum (i.e., short data-window). Whereas $\widetilde{\rho} \approx 1$ during sustained voicing, ρ seems to saturate at a lower value. Thus, using ρ and $\widetilde{\rho}$ directly in (6) would bias the decision in favour of the long window.

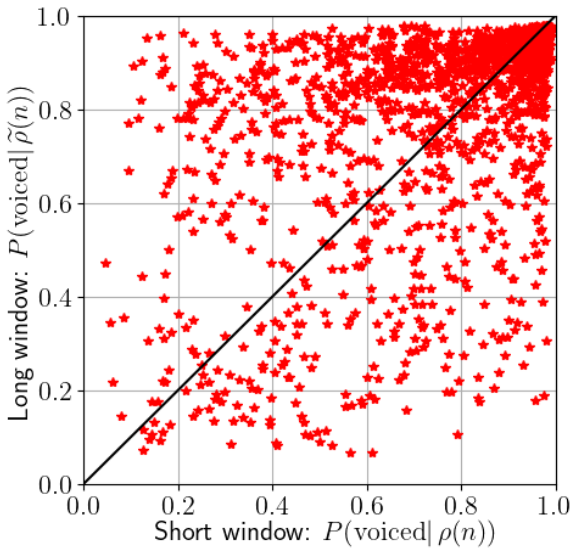


Fig. 4. Comparison of the voicing probabilities using the logit models for the normal ($\mathbf{Y}$) and the refined ($\widetilde{\mathbf{Y}}$) spectra. The distribution now shows no bias towards one or the other approach and the reliability indicators reach approximately 1 in both cases. The voicing probability, thus derived, may now be used in (6) to effect a choice between using the the original spectrum or the refined spectrum for voicing detection and F_0 estimation.

that the data-point $\mathbf{x}$ belongs to *Class 1*. Thus, we can also write: $\mathcal{L}(\mathbf{x}) = P\big(\text{Class } 1\big|\mathbf{x}\big)$.

In our case, the input feature is $\rho(n)$ (resp. $\widetilde{\rho}(n)$). The aim is to determine whether the data frame that produced this value of $\rho(n)$ (resp. $\widetilde{\rho}(n)$) is *voiced* (Class 1) or *unvoiced* (Class 0). The logit model then has two parameters and, may be written as:

$$\mathcal{L}(\rho(n)) = P\big(\text{voiced}\big|\rho(n)\big) = \frac{1}{1+\exp\big(-w_0 - w_1\rho(n)\big)}\,, \tag{8}$$

when computed on the voicing feature of the original spectrum (short data window), and as:

$$\widetilde{\mathcal{L}}(\widetilde{\rho}(n)) = P\big(\text{voiced}\big|\widetilde{\rho}(n)\big) = \frac{1}{1+\exp\big(-\widetilde{w}_0 - \widetilde{w}_1\widetilde{\rho}(n)\big)}\,, \tag{9}$$

when using the voicing feature from the refined spectrum.

The models are individually trained using annotated data from a high-quality pitch estimation/tracking database such as [19]. Figure 4 shows the scatterplot obtained when predicting the probability of voicing using the logit model. It may be seen that the distribution is now more equitable, allowing for a direct comparison of the voicing reliability using the original and the refined spectrum. The voicing probabilities, thus derived, may now be used in (6) to effect a choice between using the the original spectrum or the refined spectrum for voicing detection and F_0 estimation and we can reformulate (6) as:

$$\text{Choose}\begin{cases}\widetilde{\mathbf{Y}}(n) \text{ if } P\big(\text{voiced}\big|\widetilde{\rho}(n)\big) \geq P\big(\text{voiced}\big|\rho(n)\big)\,,\\ \mathbf{Y}(n) \text{ otherwise}\,.\end{cases} \tag{10}$$

Further, a *common* threshold may now be set on the probabilities to avoid false voicing detections. This threshold can be set as an application dependent trade-off between acceptable false-alarms and missed detections.

VI. Experimental evaluation

To evaluate the benefit of the proposed adaptive switching within the spectral refinement framework, we use the PTDB-TUG database [19] – a high-quality audio database, with reliable annotations, developed for testing and evaluating pitch estimation algorithms. The sampling frequency used is $F_s = 16\,\text{kHz}$ and the spectral analysis parameters are fixed to $N = 512$ with a frameshift of 256 samples. The signals are windowed by a periodic Hann window before the DFT.

The spectral refinement matrix is generated as proposed in [11] and applied to the short-term spectrum as described in II. To enable a fair comparison, the SR matrix is designed such that the refined spectrum is computed for the *same* frequencies as the original spectrum. However, effectively $\widetilde{N} = 1024$ data points are considered which, in conjunction with the chosen frameshift of 256 samples leads to a memory of $L = 3$ frames in the SR matrix.

The parameters of the logistic regression classifier are trained on a gender-balanced subset of 20 (clean) sentences from male and female speakers. These parameters are subsequently fixed for the further analysis. The voicing detection threshold for the logit outputs was fixed at 0.6, yielding a 10 % false-alarm rate on the training data.

The evaluation is carried out on a total of 500 sentences (gender-balanced) on clean as well as corrupted conditions. Since one application area of this research is automotive, for the noise we have selected the car and traffic noises from the

		Original spectrum ($\mathbf{Y}$)	Refined spectrum ($\widetilde{\mathbf{Y}}$)	Adaptive switching
Clean	FA	11.0 %	10.9 %	15.9 %
	TP	80.9 %	85.2 %	91.6 %
0 dB	FA	11.0 %	13.89 %	20.4 %
	TP	72.6 %	81.0 %	88.27 %
10 dB	FA	12.6 %	13.2 %	19.4 %
	TP	80.5 %	85.3 %	91.8 %

TABLE I
VOICING DETECTION PERFORMANCE USING THE ORIGINAL SPECTRUM, THE REFINED SPECTRUM AND THE PROPOSED ADAPTIVE SWITCHING BETWEEN THE TWO.

		Original spectrum ($\mathbf{Y}$)	Refined spectrum ($\widetilde{\mathbf{Y}}$)	Adaptive switching
Clean	GR(%)	4.0 %	3.4 %	2.7 %
	FE(Hz)	2.1	3.72	3.57
0dB	GR(%)	10 %	12.4 %	9.1 %
	FE(Hz)	2.4	3.9	3.73
10dB	GR(%)	5.7 %	4.9 %	3.7 %
	FE(Hz)	2.2	3.8	3.6

TABLE II
F_0 ESTIMATION ACCURACY WHEN USING THE ORIGINAL SPECTRUM, THE REFINED SPECTRUM AND THE PROPOSED ADAPTIVE SWITCHING BETWEEN THE TWO.

ETSI database [20] and consider mixing conditions of 0 dB and 10 dB signal-to-noise ratios. The performance is evaluated in terms of the voicing detection and in terms of the F_0 estimation error.

A. Voicing detection

For the voicing detection accuracy, the false-alarm (FA) and true-positive (TP) rates are presented. False-alarm means that the algorithm under test predicts voicing where no reference pitch is available. The results are summarised in Table. I.

B. F_0 estimation accuracy

The voicing detection performance indicates the capability of detecting voicing, independent of the actual F_0 estimate. Metrics evaluating F_0 estimation accuracy, therefore, benchmark performance along an orthogonal dimension and, together, can give a more complete picture of the algorithms' potential. The accuracy is typically described using two metrics: the *gross error rate* and the *fine error*.

The gross error rate indicates how often the pitch estimation algorithm produces a result that is *widely different* from the ground truth. The threshold for indicating an error as a gross error can be an absolute value or can be specified as a percentage of the deviation from the ground truth. We opt for the former definition for the following reason: since we operate in the discrete frequency domain, pitch estimates that resolve to the same discrete frequency bin as the ground truth will, in general, be accurate enough for most applications (speech enhancement, phase restoration). Thus, defining a gross error as a percentage of the ground truth will unfairly penalise estimation errors for low F_0, even though the effect of such errors in a practical application would be negligible. In our case, since the frequency resolution for the chosen DFT size ($N = 512$) is ≈ 30Hz, we choose this value to define the gross error threshold. This way, frames with gross error resolve the estimated F_0 to a discrete frequency at least one bin removed from the true value – which strongly impacts further processing that depends on the pitch estimate. We believe this provides a more indicative result. The gross-error rate (GR) is, therefore, presented as the percentage of detected voiced frames where the F_0 estimation error is larger than the set threshold:

$$GR = \frac{\sum_n I_n}{\text{total voiced frames}}\,, \qquad (11)$$
$$\text{where } I_n = \begin{cases} 1 & |\widehat{F_0}(n) - F_0(n)| > 30\text{Hz}\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

and $\widehat{F_0}(n)$ is the estimated fundamental frequency by any method whereas $F_0(n)$ is the ground truth value.

The fine error (FE) is the average error in the F_0 estimate for frames where the error is within the gross-error threshold. This metric then indicates the average accuracy of the estimation error in frames where the estimate is close to the ground truth:

$$FE = \frac{\sum_n |\widehat{F_0}(n) - F_0(n)| J_n}{\sum_n J_n}\,, \qquad (12)$$
$$\text{where } J_n = \begin{cases} 1 & |\widehat{F_0}(n) - F_0(n)| \leq 30\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

The results are summarised in Table II.

VII. DISCUSSION

The experimental results for the voicing detection on *clean speech* indicate that always using the refined spectrum can provide a better detection. This is in line with the previous research on SR. However, that there is still potential for improvement is indicated by the results when using the proposed adaptive switching scheme – which offers a significant improvement in the true-positive rate with an acceptable increase in the false-alarm rate.

In noisy conditions, the improvement in the positive detection rate offered by the refined spectrum and the proposed adaptive switching is even more marked compared to the original spectrum, at the cost of a small increase in the false-alarm rate. The increase in FA is, however, more than compensated by the improvement in the true positive rate for these situations.

The benefit of the proposed adaptive switching scheme, however, can best be appreciated on the F_0 estimation accuracy. Whereas in clean and high SNR conditions, using the refined spectrum leads to a lower gross-error rate compared to using the original spectrum, the performance degrades in low SNR conditions. This confirms our hypothesis that the use of the refined spectrum (with the correspondingly longer data window) may not always be beneficial. Switching between the original and the refined spectrum in the proposed adaptive manner yields the *lowest gross error rate* for all conditions!

The fine error for all approaches is within the same order of magnitude. The slightly higher fine error when using the refined spectrum is to be expected. The refined spectrum effectively uses a longer data window and thus averages the periodicity of the voiced segment over a larger interval, which smooths out the estimate for segments with quickly varying

F_0. The original spectrum, which uses a shorter window, better captures these dynamics and leads, consequently, to a smaller FE. The proposed method, with the adaptive switching, has an FE that is in between these two, as expected.

VIII. Conclusions

Spectral refinement is a computationally inexpensive way to improve spectral resolution by linearly combining successive short-term spectra. Thereby we can base the spectral analysis for subsequent processing on a long data window, without compromising on latency and without significant computational overhead. This has been previously applied to the problem of F_0 estimation and has been shown to be beneficial especially in the estimation of very low F_0. However, using the refined spectrum leads to poorer detection of voicing onsets (due to the Heisenberg-Gabor limit) and a poorer performance in low SNR conditions. Switching adaptively between the original short-term spectrum and the refined version provides a good trade-off in these situations.

The switching logic is based on the reliability of voicing which, in turn, is based on the maximum of the normalised auto-correlation within the range of time-lags corresponding to expected periodicity for speech. However, it was demonstrated that this feature cannot be used *directly* for comparing the estimates from the refined and original spectra, since it would bias the decision towards the refined spectrum. Hence, we proposed the use of logistic regression classifiers, individually trained for each spectrum, to *transform* the voicing feature into a *probability of voicing*. These probabilities are now directly comparable and can be used to switch between the F_0 estimates from the refined and original short-term spectra.

The benefit of the proposed adaptive switching was conclusively demonstrated on speech data in a variety of clean and corrupted conditions. In terms of voicing detection performance, adaptive switching provides a significant improvement in the true-positive rate, even in low SNR conditions, compared to using either of the individual spectra exclusively. This improvement comes at the cost of a small increase in the false-alarm rate, which may be deemed acceptable. In terms of the F_0 estimation accuracy, the adaptive switching method provides the best result in terms of gross-error rate while the fine error is of the same order of magnitude for all approaches.

Lastly, we note that while we have demonstrated the results on a specific F_0 estimator, the discussion and results retain their validity for most state-of-the-art F_0 estimators, since these are (implicitly or explicitly) based on the short-term spectra of the signals.

IX. Acknowledgements

The authors would like to express their sincere thanks to Mr. Zeno Verboven (UGent) and Mr. Robin Hick (TH Aschaffenburg) for their assistance with coding the spectral refinement approach and for improving the annotations in the pitch database. The work of Nilesh Madhu is supported by a research grant from UGent (BOF.STG.2019.0008.01).

References

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control – A Practical Approach*. Hoboken, NJ, USA: John Wiley & Sons, 2004.

[2] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Hoboken, NJ, USA: John Wiley & Sons, 2006.

[3] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process*, vol. 14, no. 6, pp. 2098–2108, 2006.

[4] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Two-stage speech enhancement with manipulation of the cepstral excitation," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, March 2017, pp. 106–110.

[5] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Dnn-supported speech enhancement with cepstral estimation of both excitation and envelope," *IEEE Trans. Audio, Speech, and Language Process*, vol. 25, no. 8, pp. 1592–1605, 2018.

[6] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, Dec 2014.

[7] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 124, pp. 1638–1652, 2008.

[8] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Elsevier Signal Processing*, vol. 135, p. 188–197, 2017.

[9] ETSI, "Ets 300 903 (GSM 03.50): Transmission planning aspects of the speech service in the GSM public land mobile network (PLMS) system," 1999.

[10] G. Schmidt and T. Haulick, "Signal processing for in-car communication systems," *Signal Processing*, vol. 86, no. 6, pp. 1307–1326, 2006.

[11] M. Krini and G. Schmidt, "Spectral refinement and its application to fundamental frequency estimation," in *Proc. Workshop on applications of signal processing to audio and acoustics (WASPAA)*, 2007.

[12] M. Krini and N. Madhu, "Generalized refinement of short-term fourier spectra in time-and frequency-domain and its combination with polyphase filterbanks," in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec 2019.

[13] ——, "Improved F_0 estimation by generalised spectral refinement applied to dft-modulated polyphase filterbanks," in *2019 Ninth International Symposium on Embedded Computing and System Design (ISED)*, Dec 2019, pp. 1–5.

[14] A. Mertins, *Signal Analysis*. Chichester, England: John Wiley & Sons, 1999.

[15] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Trans. Speech and Audio Process*, vol. 7, no. 3, pp. 333–338, 1999.

[16] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[17] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[18] J. S. Cramer, "The origins of logistic regression," Tinbergen Institute Working Paper No. 2002-119/4, Tech. Rep., 2002. [Online]. Available: doi:10.2139/ssrn.360300

[19] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. INTERSPEECH*, 2011, pp. 1509–1512.

[20] ETSI, "Eg 202 396-1: Speech processing, transmission and quality aspects (stq); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database, european telecommunications standards institute," Sep. 2008.