A Spectral Variation Function for Variable Time-Scale Modification of Speech

Pramod H. Kachare 1, 2, Prem C. Pandey 1

1 Dept. of Electrical Engg., Indian Institute of Technology Bombay, Mumbai, India
2 Dept. of Electronics & Telecom. Engg., Ramrao Adik Institute of Technology, Navi Mumbai, India

kachare.pramod1991@gmail.com, pcpandey@ee.iitb.ac.in

Abstract— Spectral variation function is used to detect salient segments (segments with sharp spectral transitions). It is calculated from cosine of the angle between the averaged feature vectors of the adjacent segments. A modified version of this function is presented for variable time-scale modification of the speech signal. It uses the magnitude spectrum smoothed by auditory critical band filters and a small offset in the normalization for the angle cosine. Test results showed that the modified function detects spectral saliencies and does not have spurious peaks. It is suited for variable time-scale modification without altering the overall duration. Listening tests showed significantly better speech quality for processing using the modified function.

Keywords—Spectral variation function, time-scale modification, voice conversion

I. INTRODUCTION

Time-scale modification (TSM) of the speech signal is used for its compression or expansion. A time-expanded signal can improve speech intelligibility for hearing and language impaired children [1]–[3]. Its use in digital playback can help foreign-language speakers and the elderly to improve speech comprehension [4]. Prosody transformation in voice conversion lengthens or shortens the source duration to match the target duration [5]–[7]. A uniform TSM reduces the timing differences between the source and the target duration patterns, but it is generally insufficient for high-quality voice conversion [8]. For large time-scaling factors, uniform TSM results in loss of speech intelligibility due to disruption of transient acoustic cues in the consonants [8].

The TSM techniques for compression or expansion of the speech signal with a uniform time-scaling factor are mostly based on synchronous overlap-add (SOLA) or phase vocoder approaches [19]–[26]. The variable TSM techniques generally segment the speech signal into different segments and use a SOLA variant with a variable time-scaling factor based on the segment type and the desired average time-scaling factor [9]–[16].

The method of Lee et al. [9] detects the transient segments using the LPC cepstral distance between neighbouring frames and uses a modified SOLA method to retain the transient segments and to compress or expand the steady-state segments. The method of Covell et al. [10] compresses the speech signal to mimic the natural fast speech. It calculates 'audio tension' as a function of normalized frame energy and normalized first difference of the short-time magnitude spectrum and uses a modified SOLA a method with lower compression for segments with higher tension. The method of Pesce [11] stretches the speech signal using a SOLA variant, retaining the transients and expanding the steady-state segments. It detects the transients using zero crossing rate and rate of change of the short-time energy. The method of Demol et al. [12] uses the average magnitude difference function and the normalized average magnitude difference in loss of speech intelligibility due to disruption of transient acoustic cues in the consonants [8].

Several investigations have used spectral variation function (SVF) to detect salient segments (segments with sharp spectral transitions) for use in speech recognition [28]–[33]. A modified version of this function is presented to obtain the time-scaling factor, without explicitly marking the segment classes, for variable TSM using single-pass processing. The proposed function is used in a variable TSM framework based on 'SOLA with fixed synthesis' (SOLAFS) [19]. The second section describes the proposed SVF, and the test results are given in the third section. Application of the proposed function for variable time-scale speech modification is presented in the fourth section, followed by the conclusion in the last section.

II. PROPOSED SPECTRAL VARIATION FUNCTION

The SVF provides a measure of discrepancy in the mean normalized spectra of two adjacent signal segments. It is
calculated from the angle between the mean feature vectors representing the two segments. It has a low value if the two segments have no spectral variation and a high value if there is a spectral saliency marked by sharp spectral transition. Several studies have used SVFs based on different acoustic features to detect spectral saliencies for speech segmentation [28]-[33]. In these studies, the SVF is calculated as a function of the frame position \( n \), using \( L \) frames on either side. Let the \( n \)-th frame features, with feature index \( k \), be \( \{X(n, k), 1 \leq k \leq K \} \). The 2\( L+1 \) frames centered on the \( n \)-th frame are used to calculate the mean-subtracted averaged features for the left and the right sides as the following:

\[
X_l(n, k) = \frac{1}{L} \left( \sum_{i=-L}^{L} X(i, k) - \frac{1}{2} \sum_{i=-L}^{L} X(i, k) \right) \quad (1)
\]

\[
X_r(n, k) = \frac{1}{L} \left( \sum_{i=-L}^{L} X(i, k) - \frac{1}{2} \sum_{i=-L}^{L} X(i, k) \right) \quad (2)
\]

The cosine of the angle between the feature vectors \( X_l(n) \) and \( X_r(n) \) is calculated as their normalized inner product as

\[
\rho(n) = \frac{X_l(n) \cdot X_r(n)}{\left( \|X_l(n)\| \|X_r(n)\| \right)^{1/2}} \quad (3)
\]

It may be noted that division by \( L \) in (1) and (2) is not needed for the angle cosine calculation in (3). The SVF as defined by Esposito and Aversano [31], denoted as \( F_{EA} \), is calculated from the angle cosine as

\[
F_{EA}(n) = (1 - \rho(n))/2 \quad (4)
\]

This function’s range is \([0, 1]\) independent of the number and range of the features, with zero representing no spectral transition and one representing a sharp transition.

The SVFs calculated using different features have been used for speech segmentation in HMM-based automatic speech recognition on TI/NIST connected digit corpus [28], [29] and TIMIT corpus [30], [31]. Esposito and Aversano [31] evaluated the functions based on linear-frequency cepstrum, mel-frequency cepstrum, and mel filterbank outputs and reported best performance for the mel filterbank outputs. Application of these functions for variable TSM of the speech signal results in perceptible distortion. A detailed examination of the input speech signals, SVFs, and the time-scale modified signals showed the distortion to be related to the presence of spurious peaks in the SVF. These spurious peaks were observed to be related to low-energy segments and variations in the high-frequency part of the spectrum.

Several modifications for suppressing the spurious peaks in the SVF without affecting its relationship with the spectral saliencies were investigated. The SVF calculation using the short-time spectrum represented by auditory critical band filters or outputs of a mel filterbank de-emphasizes the variation in the high-frequency part of the spectrum. Thus, it should help detect the spectral saliencies of perceptual significance. However, the SVF calculated from such a representation gets affected by the band center frequencies because a spectral variation near the band edges contributes more to the SVF than near the band centers. Therefore, it is proposed to calculate the SVF using the magnitude spectrum smoothed by auditory critical band filters centered at each frequency sample and inversely weighted by the filter bandwidth. It de-emphasizes variation in the high-frequency part of the spectrum without downsampling the spectrum as a set of discrete bands. For suppressing spurious peaks during the low-energy segments, it is proposed to introduce a small offset in the normalization for calculating the cosine of the angle between the left and the right mean-subtracted average feature vectors.

Let the short-time magnitude spectrum of the input speech signal, with sampling frequency \( f_s \), be calculated using \( M \)-point window, \( M \)-point window shift, and \( N \)-point DFT, with \( M < M \) and \( N > M \). Let the magnitude spectrum for the \( n \)-th frame be \( \{S(n, k), 0 \leq k \leq N/2 \} \). The magnitude spectrum is smoothed using auditory critical band filters. The auditory critical bandwidth as a function of the frequency is calculated as in [34]. The bandwidth of the filter centered at the frequency index \( k \) is calculated as

\[
BW(k) = \frac{N}{f_s} \left[ 25 + 75 \left( 1 + 1.4 \left( \frac{k - f_s}{N} \right)^{0.59} \right) \right] \quad (5)
\]

The low and high ends of this band are calculated as the following:

\[
a(k) = k - \left \lfloor \frac{BW(k)/2} \right \rfloor \quad (6)
\]

\[
b(k) = k + \left \lceil \frac{BW(k)/2} \right \rceil \quad (7)
\]

The filter used has a triangular magnitude response with a peak at \( k \) and given as

\[
H_s(m) = \begin{cases} 
2(m-a(k))/(b(k)-a(k)), & a(k) < m < k \\
2(b(k)-m)/(b(k)-a(k)), & k \leq m < b(k) \\
0, & \text{otherwise}
\end{cases} \quad (8)
\]

The smoothed magnitude spectrum \( X(n, k) \) is calculated as a correlation of \( S(n, m) \) with \( H_s(m) \) and inversely weighted by the filter bandwidth, and it is given as

\[
X(n, k) = \sum_{m=a(k)}^{b(k)} S(n, m)H_s(m) \left/ BW(k) \right. \quad (9)
\]

The mean-subtracted averaged vectors \( X_l(n) \) and \( X_r(n) \) are calculated as in (1) and (2), respectively. Low signal energy during the unvoiced and silence segments can cause numerical instability in ratio calculation in (3). For avoiding this problem, a small offset \( \epsilon \) is added to the normalizing factor. Thus the proposed equation for calculating the cosine of the angle between \( X_l(n) \) and \( X_r(n) \) is given as

\[
\rho_{\epsilon}(n) = \frac{X_l(n) \cdot X_r(n)}{\left( \|X_l(n)\| \|X_r(n)\| \right)^{1/2} + \epsilon} \quad (10)
\]

The offset \( \epsilon \) is set as a fraction of the maximum of the magnitude spectrum as

\[
\epsilon = 10^{-\beta} \max(S(n, k))^2 \quad (11)
\]
with an empirically selected exponent $\beta$. A 5-point median filter is applied on $\rho_s(n)$ to suppress low-level spurious ripples without significantly distorting the large saliency related variations. The resulting output is used to calculate SVF as in (4). Thus the proposed SVF, denoted as $F_p$, is given as

$$F_p(n) = \left[1 - 5\text{-point\_median}\left\{\rho_s(n)\right\}\right]/2 \quad (12)$$

The window length $M$, window shift $M_s$, exponent $\beta$, and number of frames $L$ for calculating the feature vectors are selected based on empirical investigation as described in the next section.

### III. TEST RESULTS

Calculation of the proposed SVF for different values of $M$, $M_s$, $\beta$, and $L$ was examined using sentences in the TIMIT database as the test material with the sampling frequency $f_s$ as 16 kHz. The window length was varied over 5–40 ms ($M$: 80–640 samples) with window shift of 1.25–20 ms ($M_s$: 20–320 samples). A small window resulted in a large number of spurious peaks indicating falsely detected spectral saliencies, and the number of these peaks decreased with an increase in the window length. A very large window resulted in mis-detection of spectral saliencies. A 10-ms window ($M = 160$) with a 5-ms window shift ($M_s = 80$) was found to be most suitable. The number of frames $L$ for SVF calculation was varied over 1–10, with high $L$ resulting in misdetection of spectral saliencies, low $L$ resulting in spurious peaks, and $L$ as 2 to be optimal. This combination of $f_s$, $M$, $M_s$, and $L$ corresponds to using 55-ms segments for SVF calculation. The floor parameter $\beta$ was varied over 1–10, and $\beta$ as 6 provided a consistent detection of spectral saliencies.

The proposed SVF was calculated for signals with sampling frequency $f_s$ as 16 kHz using $M = 160$, $M_s = 80$, $N = 320$, $\beta = 6$, and $L = 2$. For comparison, the SVF based on mel filterbank outputs as described by Esposito and Aversano [28] was calculated using eight mel filterbank outputs over 0–8 kHz bandwidth, $M = 160$, $M_s = 80$, and $L = 4$. An example of the SVF calculation using both methods is shown in Fig. 1 for the TIMIT sentence "She had your dark suit in greasy wash water all year" spoken by a female speaker. It can be observed that the earlier reported SVF shows spurious peaks during the stop closures and unvoiced fricatives, while the proposed SVF marks the spectral saliencies by a large value without having spurious peaks. Similar results were obtained for other sentences.

### IV. TIME-SCALE MODIFICATION USING THE PROPOSED SVF

Variable TSM of the audio signal is implemented using the modified SOLAFS technique [19] with a variable time-scaling factor and single-pass processing. It expands the salient segments and compresses other segments by an appropriate time-scaling factor, preserving the overall duration. The time-scaling factor $\alpha(n)$ for the $n$th frame is obtained from the SVF $F(n)$ using a piecewise linear mapping as the following:

$$\alpha(n) = \begin{cases} \alpha_{\text{min}}, & F(n) \leq F_{\text{min}} \\ \alpha_{\text{max}} + (\alpha_{\text{max}} - \alpha_{\text{min}}) \frac{F(n) - F_{\text{min}}}{F_{\text{max}} - F_{\text{min}}}, & F_{\text{min}} < F(n) < F_{\text{max}} \\ \alpha_{\text{min}}, & F_{\text{max}} \leq F(n) \end{cases} \quad (13)$$

Here $F_{\text{min}}$ and $F_{\text{max}}$ are the lower and upper limits of $F(n)$, respectively, and $\alpha_{\text{min}}$ and $\alpha_{\text{max}}$ are the corresponding limits of $\alpha(n)$. The mapping is shown in Fig. 2. For an average time-scaling factor $\overline{\alpha}$, $\alpha_{\text{min}}$ is calculated as

$$\alpha_{\text{min}} = \max\left(0, \left(\overline{\alpha} - \alpha_{\text{max}}\right) \frac{F_{\text{max}} - F_{\text{min}}}{F_{\text{max}} - F_{\text{min}}} \left(1 - \frac{F_{\text{max}} - F_{\text{min}}}{F_{\text{max}} - F_{\text{min}}}\right)\right) \quad (14)$$

where $F$ is the mean of the $F(n)$ values between $F_{\text{min}}$ and $F_{\text{max}}$. If either $\alpha_{\text{min}}$ or $\alpha_{\text{max}}$ is equal to $\overline{\alpha}$, then (13) results in a uniform time-scaling factor. For $\overline{\alpha} = 1$, the relation permits compression or expansion of transient and steady-state...
segments of the speech signal without altering the speech signal duration.

An example of the variable TSM using the proposed SVF (F_P) and the earlier reported SVF (F_EA), for the sentence in Fig. 1, is shown in Fig. 2. The time-scaling factor for TSM was obtained using the piecewise linear mapping as in (13) and (14) and with $\alpha = 1$, $\alpha_{\text{max}} = 2$, $F_{\text{min}} = 0$, $F_{\text{max}} = 1$. Informal listening of this and other sentences indicated that the TSM signal using the earlier SVF had perceptible distortions, which were not present using the proposed SVF. Both signals had no detectable loss of intelligibility.

The difference in speech quality of variable TSM using the proposed SVF (F_P) and the earlier reported SVF (F_EA) was assessed by a mean opinion score (MOS) test [35]. The test was conducted using sentences from the Hindi Speech Database [36], comprising 500 sentences spoken by 50 speakers. Ten sentences from two speakers (a male and a female), resulting in 20 test sentences, were used as the speech material. Processing was carried out with $\alpha$ as 1 (unaltered overall speech duration) and $\alpha_{\text{max}}$ as 1.25, 1.5, 1.75, and 2. Each presentation comprised a sentence (unprocessed signal, 0.5-s silence, and processed signal). After each presentation, the listener scored the quality of the processed signal on 1–5 scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). With 20 sentences, two SVFs, and four values of $\alpha_{\text{max}}$, there were 160 presentations for each listener. The SVF and $\alpha_{\text{max}}$ combinations were randomized across the presentations. The test was conducted using a PC-based setup with a graphical user interface (GUI) for signal presentation and response recording, as shown in Fig. 3. There was a familiarization session before the actual test. The signals were presented over headphones at the most comfortable level as set by the listener. Seven listeners with normal hearing participated in the test.

Results from the MOS test are given in Table I. For $\alpha_{\text{max}} = 1$, the two SVFs result in same score. At higher values of $\alpha_{\text{max}}$, the proposed SVF (F_P) resulted in higher scores, with the score difference of 1.49, 1.41, and 1.15 for $\alpha_{\text{max}}$ of 1.50, 1.75, and 2.0, respectively. All these differences were statistically significant ($p < 1\%$). Thus the results of the speech quality assessment indicate better suitability of the proposed SVF for variable TSM.

### V. CONCLUSION

A modified spectral variation function has been presented for variable TSM of the speech signal. It uses the magnitude spectrum smoothed by auditory critical band filters and a small offset in the normalization for the angle cosine. Test results using the TIMIT sentences showed that the modified function detects spectral saliencies and does not have


35. ITU-T Rec. P.800.2 Mean opinion score interpretation and reporting, 2016.