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 applied 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]. Non-uniform or variable TSM can provide a high-quality speech without intelligibility degradation [9]–[16]. It can be used to compress audio for synchronization in the audio-visual speed-up [17], [18]. It can also be used to improve speech perception under adverse listening conditions by expanding the transient consonantal segments and correspondingly compressing the steady-state segments without altering the overall duration [13], [14].

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 segment types 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 first difference of the short-time energy to mark each speech segment as a pause, vowel-like, consonant-like, phone transition, or plosive-like segment. It modifies the speech to correspond to natural slow or fast speech using the waveform-SOLA (WSOLA) method [20] with a time-scaling factor that depends on the segment type and the desired average time-scaling factor. In the method of Jayan *et al.* [13], the transients are detected using the normalized product of the mean of the first difference of the spectral centroid and energy of five bands. These segments are expanded with a corresponding compression of the steady-state segments to maintain the overall signal duration. he method of Jayan and Pandey [14], the transients are detected using GMM parameters of the short-time log-magnitude spectrum. The methods in [13] and [14] use harmonic-plus-noise model to generate the time-scaled speech signal. The method of Grofit and Lavner [15] uses a modified WSOLA method to expand the steady-state segments and retain the transients, which are detected using the L2-norm of MFCC difference and the normalized cross-correlation between consecutive frames. The method of Kupryjanow and Czyzewski [16] detects the consonants using 'peak-valley difference' calculated using a pre-trained vowel model as in [27] and uses a SOLA variant to stretch the speech signal by expanding the vowels more than the consonants.

The methods for marking the speech segments for estimating the variable time-scaling factor in the above techniques are generally computation-intensive and often not suited for single-pass processing in real-time applications. 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

978-1-6654-4177-3/21/$31.00 ©2021 IEEE

978-1-6654-4177-3/21/$31.00 ©2021 IEEE

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 nth frame features, with feature index k, be $\{X(n, k), 1 \le k \le K\}$. The $2L$+1 frames centered on the nth frame are used to calculate the mean-subtracted averaged features for the left and the right sides as the following:

$$\bar{X}_l(n, k) = \frac{1}{L}\left(\sum_{i=n-L}^{n} X(i, k) - \frac{1}{2}\sum_{i=n-L}^{n+L} X(i, k)\right) \qquad (1)$$

$$\bar{X}_r(n, k) = \frac{1}{L}\left(\sum_{i=n}^{n+L} X(i, k) - \frac{1}{2}\sum_{i=n-L}^{n+L} X(i, k)\right) \qquad (2)$$

The cosine of the angle between the feature vectors $\bar{\mathbf{X}}_l(n)$ and $\bar{\mathbf{X}}_r(n)$ is calculated as their normalized inner product as

$$\rho(n) = \frac{\bar{\mathbf{X}}_l(n) \cdot \bar{\mathbf{X}}_r(n)}{\left[\left(\bar{\mathbf{X}}_l(n) \cdot \bar{\mathbf{X}}_l(n)\right)\left(\bar{\mathbf{X}}_r(n) \cdot \bar{\mathbf{X}}_r(n)\right)\right]^{1/2}} \qquad (3)$$

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

$$F_{EA}(n) = (1 - \rho(n))/2 \qquad (4)$$

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 bands 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_s-point window shift, and N-point DFT, with $M_s < M$ and $N > M$. Let the magnitude spectrum for the nth frame be $\{S(n, k), 0 \le k \le N/2\}$. 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

$$\mathrm{BW}(k) = \left\lfloor \frac{N}{f_s} \left\lfloor 25 + 75\left(1 + 1.4\left(\frac{k f_s}{N}\right)^2\right)^{0.69} \right\rfloor \right\rfloor \qquad (5)$$

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

$$a(k) = k - \lfloor \mathrm{BW}(k)/2 \rfloor \qquad (6)$$

$$b(k) = k + \lfloor \mathrm{BW}(k)/2 \rfloor \qquad (7)$$

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

$$H_k(m) = \begin{cases} 2(m - a(k))/\left(b(k) - a(k)\right), & a(k) < m < k \\ 2(b(k) - m)/\left(b(k) - a(k)\right), & k \le m < b(k) \\ 0, & \text{otherwise} \end{cases} \qquad (8)$$

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

$$X(n,k) = \left[\sum_{m=a(k)}^{b(k)} S(n,m) H_k(m)\right] \Big/ \mathrm{BW}(k) \qquad (9)$$

The mean-subtracted averaged vectors $\bar{\mathbf{X}}_l(n)$ and $\bar{\mathbf{X}}_r(n)$ 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 ε is added to the normalizing factor. Thus the proposed equation for calculating the cosine of the angle between $\bar{\mathbf{X}}_l(n)$ and $\bar{\mathbf{X}}_r(n)$ is given as

$$\rho_P(n) = \frac{\bar{\mathbf{X}}_l(n) \cdot \bar{\mathbf{X}}_r(n)}{\left[\left(\bar{\mathbf{X}}_l(n) \cdot \bar{\mathbf{X}}_l(n)\right)\left(\bar{\mathbf{X}}_r(n) \cdot \bar{\mathbf{X}}_r(n)\right)\right]^{1/2} + \varepsilon} \qquad (10)$$

The offset ε is set as a fraction of the maximum of the magnitude spectrum as

$$\varepsilon = 10^{-\beta}[\max(S(n,k))]^2 \qquad (11)$$

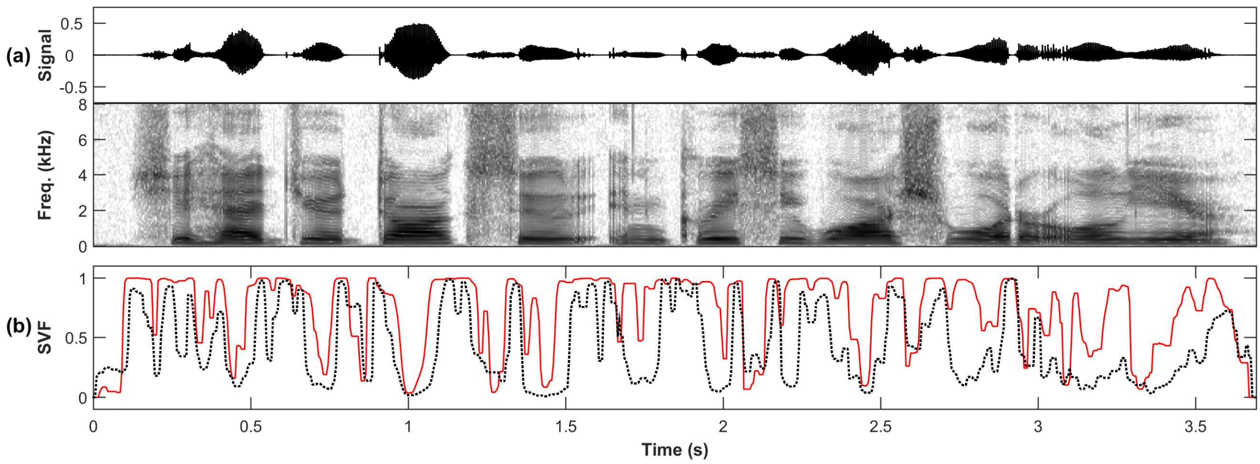

Fig. 1. SVF for TIMIT sentence "She had your dark suit in greasy wash water all year": (a) signal and its spectrogram, (b) proposed SVF (F_P, dotted black trace) and earlier reported SVF (F_{EA}, solid red trace).

with an empirically selected exponent β. A 5-point median filter is applied on $\rho_P(n)$ 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

$$F_P(n) = \left[1 - \text{5-point\_median}\left\{\rho_P(n)\right\}\right] / 2 \tag{12}$$

The window length M, window shift M_s, exponent β, 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, β, 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 β was varied over 1–10, and β 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.

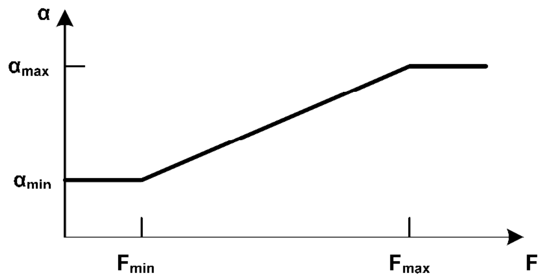

Fig. 2. Piecewise linear mapping from SVF (F) to time-scaling factor (α).

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 nth frame is obtained from the SVF $F(n)$ using a piecewise linear mapping as the following:

$$\alpha(n) = \begin{cases} \alpha_{\min}, & F(n) \le F_{\min} \\ \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \dfrac{F(n) - F_{\min}}{F_{\max} - F_{\min}}, & F_{\min} < F(n) < F_{\max} \\ \alpha_{\max}, & F_{\max} \le F(n) \end{cases} \tag{13}$$

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

$$\alpha_{\min} = \max\left(0, \left(\bar{\alpha} - \alpha_{\max} \frac{\bar{F} - F_{\min}}{F_{\max} - F_{\min}}\right) \Big/ \left(1 - \frac{\bar{F} - F_{\min}}{F_{\max} - F_{\min}}\right)\right) \tag{14}$$

where $\bar{F}$ is the mean of the $F(n)$ values between $F_{\min}$ and $F_{\max}$. If either $\alpha_{\max}$ or $\alpha_{\min}$ is equal to $\bar{\alpha}$, then (13) results in a uniform time-scaling factor. For $\bar{\alpha} = 1$, the relation permits compression or expansion of transient and steady-state

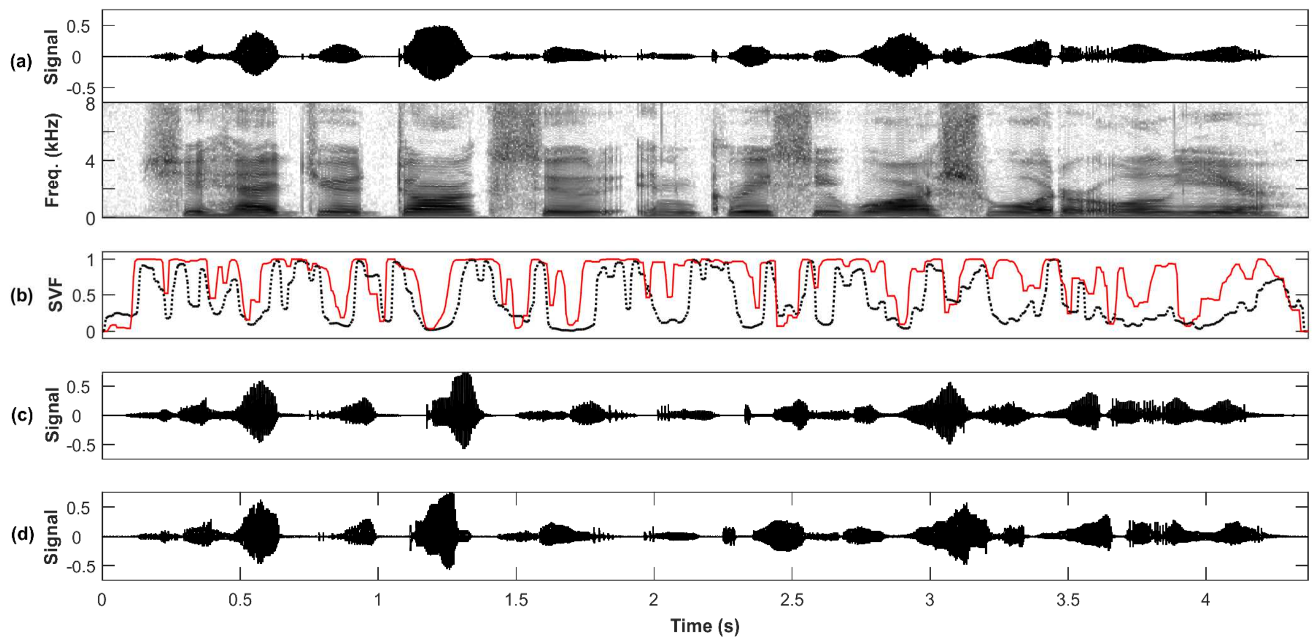

Fig. 3. An example of variable TSM for TIMIT sentence "She had your dark suit in greasy wash water all year": (a) signal and its spectrogram, (b) proposed SVF (F_P, dotted black trace) and earlier reported SVF (F_{EA}, solid red trace), (c) TSM signal using the proposed SVF (F_P), (d) TSM signal using the earlier SVF (F_{EA}).

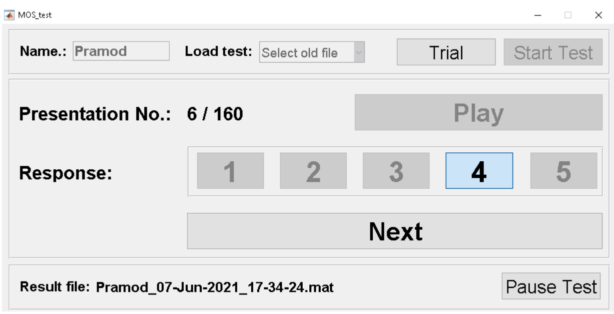

Fig. 4. GUI for MOS test.

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 $\bar{\alpha} = 1$, $\alpha_{\max} = 2$, $F_{\min} = 0$, $F_{\max} = 1$. 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 $\bar{\alpha}$ as 1 (unaltered overall speech duration) and α_{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 $\alpha_{\max}$, there were 160 presentations for each listener. The SVF and $\alpha_{\max}$ combinations were randomized across the presentations. The test was conducted using a PC-based set-up 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.

TABLE I. MEAN OPINION SCORES (MOS) FOR TSM SPEECH (NUMBER OF LISTENERS = 7, NUMBER OF PRESENTATIONS PER LISTENER=160, S.E.= STANDARD ERROR)

Max. scaling factor ($\alpha_{\max}$)	MOS: Earlier SVF (F_E) Mean	S.E.	Proposed SVF (F_P) Mean	S.E.	Difference Mean	S.E.	Paired One-tailed significance (p)
1.25	4.30	0.11	4.30	0.05	0.00	0.07	--
1.50	2.62	0.15	4.11	0.12	1.49	0.17	0.0001
1.75	2.35	0.24	3.76	0.19	1.41	0.14	0.0005
2.00	1.95	0.21	3.10	0.26	1.15	0.18	0.0005

Results from the MOS test are given in Table I. For $\alpha_{\max} = 1$, the two SVFs result in same score. At higher values of $\alpha_{\max}$, the proposed SVF (F_P) resulted in higher scores, with the score difference of 1.49, 1.41, and 1.15 for α_{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 spurious peaks. It has been applied and tested for TSM using modified SOLAFS technique and a piecewise linear mapping from the SVF to the time-scaling factor. There was no detectable loss of intelligibility. Listening tests for speech quality assessment showed variable TSM using the proposed function to be significantly better than the earlier reported function. Further work may involve its application with different SVF to time-scaling factor mappings, other TSM techniques, evaluation of variable TSM for intelligibility improvement under adverse listening conditions, and its use for voice conversion.

REFERENCES

[1] M. H. P. Stollman, T. S. Kapteyn, and B. W. Sleeswijk. "Effect of time-scale modification of speech on the speech recognition threshold in noise for hearing-impaired and language-impaired children," *Scand. Audiology,* vol. 23, no. 1, pp. 39–46, 1994.

[2] P. Tallal, S. L. Miller, G. Bedi, G. Byma, X. Wang, S. S. Narayanan, C. Schreiner, W. M. Jenkins, and M. M. Merzenich, "Language comprehension in language-learning impaired children improved with acoustically modified speech," *Sci.*, vol. 271, no. 5245, pp. 81–84, 1996.

[3] A. Kupryjanow and A. Czyżewski, "Time-scale modification of speech signals for supporting hearing impaired school children," in *Proc. Int. Conf. NTAV/SPA*, Poznan, Poland, pp. 159–162, 2009.

[4] N. J. Versfeld and W. A. Dreschler, "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners," *J. Acoust. Soc. Am.,* vol. 111, no.1, pp. 401–408, 2002.

[5] K. S. Rao and B. Yegnanarayana, "Voice conversion by prosody and vocal tract modification," in *Proc. 9th Int. Conf. Inf. Technol. (ICIT)*, Bhubaneswar, India, pp. 111–116, 2006.

[6] Y. Stylianou, "Voice transformation: A survey," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, pp. 3585–3588, 2009.

[7] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Commun.*, vol. 28, no. 3, pp. 211–226, 1999.

[8] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, vol. 80, no. 4, pp. 1016–1025, 1986.

[9] S. Lee, H. D. Kim, and H. S. Kim, "Variable time-scale modification of speech using transient information," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Munich, Germany, pp. 1319–1322, 1997.

[10] M. Covell, M. Withgott, and M. Slaney, "MACH1: nonuniform time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Seattle, WA, USA, pp. 349–352, 1998.

[11] F. Pesce, "Realtime-stretching of speech signals," *Proc. COST G-6 Conf. Digit. Audio Effects (DAFX-00)*, Verona, Italy, pp. 189–194, 2000.

[12] M. Demol, W. Verhelst, K. Struyve, and P. Verhoeve, "Efficient non-uniform time-scaling of speech with WSOLA," in *Proc. Int. Conf. Speech Comput.*, Patras, Greece, pp. 163–166, 2005.

[13] A. R. Jayan, P. C. Pandey, and P. K. Lehana, "Time-scaling of consonant-vowel transitions using harmonic plus noise model for improving speech perception by listeners with moderate sensorineural impairment," in *Proc. 19th Int. Congr. Acoust.*, Madrid, paper no. CAS-03-006, 2007.

[14] A. R. Jayan and P. C. Pandey, "Detection of stop landmarks using Gaussian mixture modeling of speech spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, pp. 4681–4684, 2009.

[15] S. Grofit and Y. Lavner, "Time-scale modification of audio signals using enhanced WSOLA with management of transients," *IEEE Trans. Audio Speech Lang. Process.,* vol. 16, no.1, pp. 106–115, 2008.

[16] A. Kupryjanow and A. Czyzewski, "A method of real-time non-uniform speech stretching," in *Proc. Int. Conf. E-Business Telecommun.* (*ICETE*), Seville, Spain, vol 314, pp. 362–372, 2011.

[17] A. Moinet, D. Thierry, and L. Philippe, "Audio time-scaling for slow motion sports videos," in *Proc. Int. Conf. Digit. Audio Effects (DAFx)*, Maynooth, Ireland, pp. 2–5. 2013.

[18] L. He and A. Gupta, "Exploring benefits of non-linear time compression," in *Proc. ACM Int. Conf. Multimedia*, Ottawa, Ontario, Canada, pp. 382–391, 2001.

[19] D. Hejna, B. R. Musicus,. "The SOLAFS time-scale modification algorithm,". *Bolt, Beranek and Newman (BBN) Tech. Rep.*, University of Cambridge, Great Britain, 1991.

[20] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Minneapolis, MN, USA, pp. 554–557, 1993.

[21] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, no. 2, pp. 175–205, 1995.

[22] P. H. W. Wong and C. Au Oscar, "Fast SOLA-based time-scale modification using modified envelope matching," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Orlando, FL, USA, vol. 3, pp. 3188–3191, 2002.

[23] D. Dorran, R. Lawlor, and E. Coyle, "High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PAOLA)," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Hong Kong, China, vol. 1, pp. 700–703, 2003.

[24] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 3, pp. 972–980, 2006.

[25] A. Moinet and T. Dutoit, "PVSOLA: a phase vocoder with synchronized overlap-add,"in *Proc. 14th Int. Conf. Digit. Audio Effects (DAFx-11)*, Paris, France, pp. 269–275, 2011.

[26] S. Rudresh, A. Vasisht, K. Vijayan, and C. S. Seelamantula, "Epoch-synchronous overlap-add (ESOLA) for time-and pitch-scale modification of speech signals," *arXiv*:1801.06492, 2018.

[27] I. C. Yoo, and D. Yook, "Robust voice activity detection using the spectral peaks of vowel sounds," *ETRI J.*, vol. 31, no. 4, pp. 451–453, 2009.

[28] F. Brugnara, A. De Mori, D. Giuliani, and M. Omologo, "Improved connected digit recognition using spectral variation functions," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Banff, Alberta, Canada, pp. 627–630, 1992.

[29] C. D. Mitchell, M. P. Harper, and L. H. Jamieson, "Using explicit segmentation to improve HMM phone recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP),* Detroit, MI, USA, pp. 229–232, 1995.

[30] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, San Diego, CA, USA, pp. 328–331, 1984.

[31] A. Esposito and G. Aversano, "Text independent methods for speech segmentation," in *Proc. Summer School Neural Netw.*, pp. 261–290, 2005.

[32] T. Svendsen and F. Soong, "On the automatic segmentation of speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, USA, pp. 77–80, 1987.

[33] G. Flammia, P. Dalsgaard, O. Andersen, and B. Lindberg, "Segment based variable frame rate speech analysis and recognition using a spectral variation function," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Banff, Alberta, Canada, pp. 983–986, 1992.

[34] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Freqenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, no. 2, p. 248, 1961.

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

[36] K. Samudravijaya, P. V. S. Rao, and S. S. Agrawal, "Hindi speech database," in *Proc. Int. Conf. Spoken Lang. Process.,* Beijing, China, 2000, pp. 456–459, 2000.