Mel-Scale Sub-band Modelling for Perceptually Improved Time-Scale Modification of Speech and Audio Signals

Neeraj Sharma, Shreepad Potadar, Srikanth Raj Chetupalli, T.V. Sreenivas
Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore-12, India.

***Abstract*—Good quality time-scale modification (TSM) of speech, and audio is a long standing challenge. The crux of the challenge is to maintain the perceptual subtilities of temporal variations in pitch and timbre even after time-scaling the signal. Widely used approaches, such as phase vocoder, and waveform overlap-add (OLA), are based on quasi-stationary assumption and the time-scaled signals have perceivable artifacts. In contrast to these approaches, we propose application of time-varying sinusoidal modeling for TSM, without any quasi-stationary assumption. The proposed model comprises of a mel-scale non-uniform bandwidth filter bank, and the instantaneous amplitude (IA), and instantaneous phase (IP) factorization of sub-band time-varying sinusoids. TSM of the signal is done by time-scaling IA, and IP in each sub-band. The lowpass nature of IA, and IP allows for time-scaling via interpolation. Formal listening tests on speech, and music (solo, and polyphonic) show reduction in TSM artifacts such as phasiness, and transient smearing. Further, the proposed approach gives improved quality in comparison to waveform synchronous OLA (WSOLA), phase vocoder with identity phase locking, and the recently proposed harmonic-percussive separation (HPS) based TSM methods. The obtained improvement in TSM quality highlights that speech analysis can benefit from appropriate choice of time-varying signal models.**

I. Introduction

Time-scale modification (TSM) of speech, and audio refers to *speeding up* or *slowing down* the playback speed without altering the perceptual attributes such as pitch, intelligibility, and naturalness. This is a classical problem [1]. A solution for TSM is useful in applications such as time alignment of multiple audio (and video) channels, personalizing listening rate in text-to-speech synthesizers, in audio interfaces for language learning, and in sonification of data.

Traditional approaches to TSM of speech, and audio are based on quasi-stationary model of the signal. Overlap-add (OLA) processing is done on short-time segments ($10-60$ ms, assuming stationarity). The popular time-domain OLA techniques include pitch-synchronous OLA (PSOLA) [2], and waveform-synchronous OLA (WSOLA) [3]. OLA processing can also be done in spectral domain via Fourier transform analysis of short-time segments, and popular techniques are phase vocoder (PV) [4], and its variant PV with identity phase locking (PV-IP) [5]. However, the quasi-stationary assumption contributes to TSM artifacts, such as transient duplication, smearing of rapid spectral variations (pertaining to transients, diphthongs, and pitch glides), and introduction of reverberation like effects (referred to as "phasiness") [5]. These artifacts lead to poor quality TSM. Variants of OLA approach have been proposed to selectively reduce the artifacts. Example, transient duplication (and also smearing) is avoided by first detecting transient regions in the signal, and then only time-shifting them [6], [7]. Recently, Driedger et al. [8] proposed decomposing the signal into two streams -namely, harmonic and percussive, using a spectrogram decomposition technique [9]. Following this decomposition PV-IP is used for TSM on the harmonic stream, and WSOLA is used for TSM on the percussive stream. The results show significant improvement in TSM quality. However, inspite of all these developments, we hypothesize that the fundamental assumption of quasi-stationarity puts a limitation on the quality achieved. A key finding of this paper is that improved quality TSM of speech, and audio can be achieved using an appropriately chosen time-varying signal model without assuming quasi-stationarity.

Early in speech research, Dudley [10] proposed speech signals as low frequency modulations riding over high frequency carriers. Later Flanagan [11] formalized this concept as phase vocoder (PV), and proposed PV as an analysis-synthesis technique. In a PV, the signal $x[n]$ is decomposed as a sum of K time-varying sinusoids as follows,

$$x[n] = \sum_{k=1}^{K} x_k[n] = \sum_{k=1}^{K} a_k[n] \cos \phi_k[n] \qquad (1)$$

where, $a_k[n]$, and $\phi_k[n]$ are the instantaneous amplitude (IA), and instantaneous phase (IP) of the k^{th} sinusoid, respectively. The first derivative of IP is referred to as the instantaneous frequency (IF). The decomposition in (1) can be obtained by passing $x[n]$ through a suitable filter bank. The IA, and IF of each filter ouptput can be computed using several ways [12], such as analytic signal formulation [13], energy separation approach [12], and signal extrema based approach [14]. For slow rate of variation in IA and IF, relative to IF itself, $x_k[n]$ is a narrowband time-varying sinusoid [15]. The crux of this paper lies in formulating TSM of $x_k[n]$ as time-scaling of IA, and IP separately, and extending it to TSM of speech, and audio signals. The implementation of the decomposition in Eqn. (1) is carried out using set of filters contiguously placed with uniform bandwidth on mel-scale [16]. The proposed approach is referred to as μTVS (**mu**lti-component **t**ime-**v**arying **s**inusoidal decomposition). We note that the sub-band

978-1-5090-5356-8/17/$31.00 © 2017 IEEE

based TSM has also been proposed by Quatieri et al. [17]. However, the filters used there are uniform bandwidth filters, and the TSM quality is found to be poor.

For comparison of the quality of TSM obtained with μTVS, the time-scaled signals are compared with short-time spectral methods such as PV, PV-IP, a short-time temporal method WSOLA, and the recent harmonic percussive separation (HPS) method [8]. It can be noted that the short-time spectral domain implementation of PV for TSM [4] is based on discrete Fourier representation. This is different from the narrowband sinusoidal decomposition in Eqn. (1); and it additionally assumes no quasi-stationary assumption. The flexibility to carry out TSM by time-scaling IA and IP signal of narrowband signals may provide reduced artifacts. The contributions of this paper are: (i) a TSM technique termed μTVS, (ii) analyzing the effectiveness of μTVS in preserving the time-varying features, and (iii) TSM quality evaluation and comparison with widely used approaches, using listening tests on a diverse set of speech, and audio signals.

II. Proposed Time-Scale Modification

The proposed approach comprises of three stages, that is analysis, modification, and synthesis, as shown in Fig. 1.

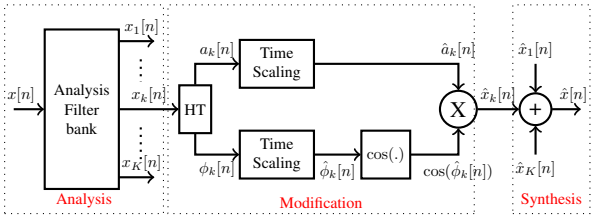

Figure 1: Block diagram of the proposed μTVS approach to time-scale modification

A. Analysis

The signal $x[n]$ is filtered through a filter bank composed of K filters. We choose a non-uniform filter bank with center frequencies spaced according to mel-scale [16], [18]. Mel-scale provides two advantages: (i) better suited for skewed spectral energy distribution in speech signals, and (ii) approximates the human auditory perception which has higher frequency resolution at lower frequencies. On the other-hand, a uniform bandwidth filter bank gives a perceptually less significant decomposition of $x[n]$. The filters are Hanning windowed sinc filters with tap length of N.

B. Modification

$x_k[n]$, the output of k^{th} channel of the filter bank, is narrowband (by construction of the filter), and we use the analytic signal approach [13] to obtain its IA, and IP. Let $x_{a,k}[n]$ be the analytic signal corresponding to $x_k[n]$. We have,

$$x_{a,k}[n] = x_k[n] + j\mathcal{H}(x_k[n]), \tag{2}$$

$$a_k[n] = |x_{a,k}[n]| \quad \text{and} \quad \phi_k[n] = \angle x_{a,k}[n] \tag{3}$$

where, $\mathcal{H}(.)$ denotes the Hilbert transform. Let α denote the time-scaling factor for TSM; $\alpha < 1$ indicates time-shrinkage, and $\alpha > 1$ indicates time-stretching. The slow temporal variations in IA, and IP of narrowband signal $x_k[n]$ simplify time-scaling. Let the input signal sampling rate be F_s. The modification of IA, and IP will benefit from oversampled IA, and IP estimates. For this, the input signal, and the filter bank are realized at a higher sampling rate (F_o). However, the filter bank is composed of K filters with uniform bandwidth along mel-scale spanning upto $F_s/2$ only. $a_k[n]$ is time-scaled by assigning the original sampled instants to corresponding time-scaled instants, and followed by interpolating $a_k[n]$ at instants corresponding to input sampling rate. Similar time-scaling is applied to unwraped $\phi_k[n]$ signal, as well. Oversampled IP provides improved unwrapping. To preserve periodicity it is important to scale IP without altering the range of variations in IF. This is done by multiplying IP by α.

Algorithm 1 μTVS method for TSM by a factor α

Over-sample: Resample $x[n]$ from $F_s(= 1/T_s)$ to $F_o(= 1/T_o)$.
Filter bank: Pass $x[n]$ through the filter bank.
IA, IP: Express $x_k[n] = a_k[n]\cos\phi_k[n]$. Estimate $a_k[n]$, and $\phi_k[n]$ (unwrapped).
Time-scaling: Assign $\hat{a}_k(t = \alpha n T_o) = a_k(nT_o)$, and $\hat{\phi}_k(t = \alpha n T_o) = \alpha\phi_k(nT_o)$.
Evaluate $\hat{a}_k(t = nT_o)$, and $\hat{\phi}_k(t = nT_o)$ using interpolation.
Express $\hat{x}_k[n] = \hat{a}_k[n]\cos\hat{\phi}_k[n]$.
Synthesis: $\hat{x}[n] = \sum_{k=1}^{K} \hat{x}_k[n]$. Resample $\hat{x}[n]$ to F_s from F_o.

C. Synthesis

At this stage we simply add the individual time-scaled narrowbands signals. The signal is down-sampled to the input sampling rate, that is F_s.

This three stage approach to TSM will be referred to as μTVS (Algorithm 1). Application of μTVS on a synthetic time-varying sinusoid is shown in Fig. 2. Both IA, and IF are time-shrinked by α. This impacts the bandwidth of $x_k[n]$ as follows. Consider $x_k(t) = a_k(t)\exp[j\phi(t)]$, and $\hat{x}_k(t) = \hat{a}_k(t)\exp[j\hat{\phi}(t)]$ as the continuous-time analytic signal corresponding to $x_k[n]$, and $\hat{x}_k[n]$, respectively. Let σ_k^2, and $\sigma_{\alpha,k}^2$, denote the bandwidth of $x_k(t)$, and $\hat{x}_k(t)$, respectively. We have [19],

$$\sigma_k^2 = \int \left(\frac{a_k'(t)}{a_k(t)}\right)^2 a_k^2(t)dt + \int \left(\phi_k'(t) - \overline{\phi_k'(t)}\right)^2 a_k^2(t)dt$$

$$\sigma_{\alpha,k}^2 = \int \left(\frac{\hat{a}_k'(t)}{\hat{a}_k(t)}\right)^2 \hat{a}_k^2(t)dt + \int \left(\hat{\phi_k}'(t) - \overline{\hat{\phi_k}'(t)}\right)^2 \hat{a}_k^2(t)dt$$

$$= \frac{1}{\alpha}\sigma_k^2, \text{ ($'$, and $\overline{(.)}$ denote derivative, and mean).} \tag{4}$$

This bandwidth scaling around the original signal IF happens in each sub-band. The dynamic range of variations in IA, and IF are same as in original. The effect of this, and other attributes of μTVS approach on speech, and audio are analyzed in Sec. III.

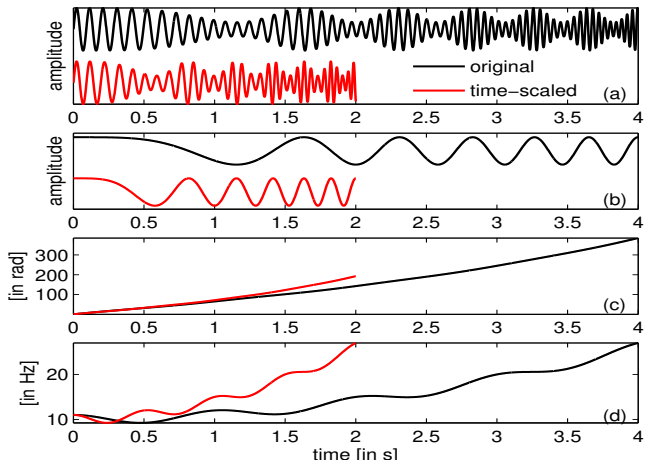

Figure 2: Illustration of TSM of a time-varying sinusoid $x_k[n] = a_k[n]\cos\phi_k[n]$. (a) $x[n]$, and time-scaled $x[n]$ (by 0.5) obtained by scaling IA, and IP. (b) IA, and time-scaled IA. (c) IP, and time-scaled IP multiplied by 0.5. (d) IF and time-scaled IF.

III. Experiments and Results

We apply the μTVS approach to speech, and music signals. All signals are sampled at $F_s = 16$ kHz. The oversampling rate for μTVS is choosen as $F_o = 96$ kHz. The filter bank parameter K is set to 32, and each filter has tap length of $N = 2048$. Since $a_k[n]$ and $\phi_k[n]$ are oversampled, piece-wise linear interpolation is found to be sufficient for the time-scaling operation. The time-scaled speech, and music signals are also compared with those obtained using PV, PV-IP, WSOLA, and HPS techniques. The implementation of these is taken from the TSM toolbox [20]. The short-time processing window duration is: 64 ms in WSOLA, 128 ms in PV, 128 ms in PV-IP, and 128 ms for harmonic stream, and 16 ms for percussion stream in HPS (close to the values in [8]). μTVS does not involve any windowing since analytic signal computation can be obtained for arbitrary large segments (of several secs).

A. Speech signals

Speech signals are composed of a range of spectrotemporal variations. Fig. 3 shows a spectrographic comparison of original, and time-scaled speech for a segment of speech sentence to highlight the effect of time-scaling on fricatives, diphthong, and plosive. For $\alpha = 0.5$, the duration of the signal is halved. In μTVS, the stationary regions, corresponding to vowels along with the periodicity (harmonics spacing), are well preserved. The time-varying spectrum corresponding to diphthongs is not preserved completely but is better than in PV-IP. Similar observations hold for regions corresponding to plosives, and fricative, as well. It should be noted that time-varying harmonics are often smeared in PV-IP approach. For $\alpha = 1.5$, duration of signal is increased by 50%. Here, the time-varying spectrum is preserved quite well for μTVS, and this is again better than PV-IP.

As discussed in (4), in μTVS approach the spectral spread in each sub-band (or channel) of the original signal gets scaled dependening on α. This is shown in Fig. 4. The spread in the

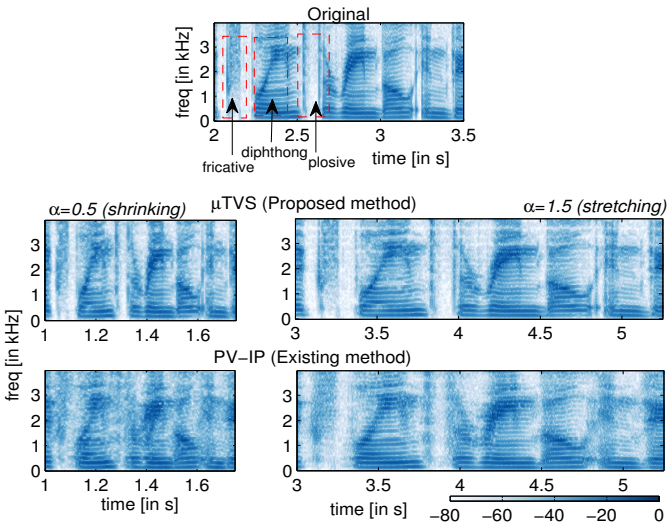

Figure 3: The original signal corresponds to the utterance: *"explained the rainbow in"*. The spectrograms (obtained with 15 ms Hanning windowed segments) of the original, and time-scaled versions are shown for two methods.

spectrum of IA of each channel gets stretched for $\alpha = 0.5$ (increase in bandwidth), and shrinks for $\alpha = 1.5$ (decrease in bandwidth). It is found that this modification of spectral spread for speech signals is perceived as natural (see Sec. III-C), and not as an artifact in the range of $0.5 < \alpha < 1.5$.

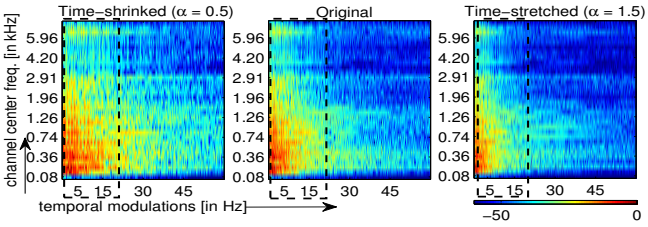

Figure 4: [In color] The Fourier spectrum (in dB) of IA signal in each of the 32 channels of the filter bank. The original signal is a 4.5 s duration female speech utterance -*"Throughout the centuries people have explained the rainbow in various ways"*.

B. Music Sound Track

Fig. 5 shows the effect of TSM on a music track composed of transient, and percussion sounds with sharp attacks, and fast (or slow) temporal decay. We see that the phase vocoder based approaches (PV, PV-IP) smear the transients, due to the quasi-stationary spectral modeling. However, PV-IP preserves the transients better than PV, as reported in [5]. The WSOLA method suffers from transient duplication due to waveform repetition (see highlighted box at the start of the track) which will affect the perceptual quality of attacks, and also of rhythm in beats. For HPS, the smearing of the transients is minimal, and also it has no transient duplication. For μTVS, the temporal structure of the sound track is preserved, and the smearing is lower than PV-IP. This shows a benefit of time-varying signal modeling in comparison to PV, and PV-

IP. Further, unlike in WSOLA, there are no transient repetition artifact.

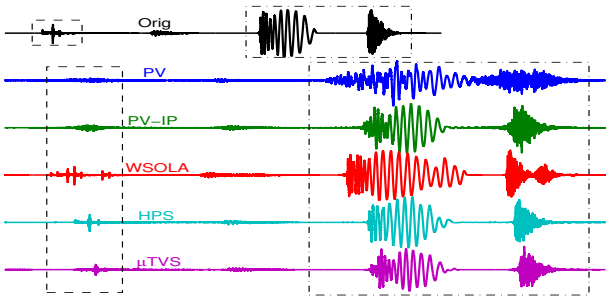

Figure 5: [In color] Effect of μTVS based TSM on a music track composed of a variety of attack and decay segments. (The individual waveforms are not aligned in time.)

C. Subjective Evaluation

To evaluate the subjective quality of the proposed TSM method μTVS a formal listening test is designed. We used seven different audio signals, chosen to span different time-varying spectral, and temporal aspects of sound. These include: *Speech* sentence (1 male, and 1 female), *Vocals* (female singing), a tune on *Flute*, a tune on *Glockenspiel* (a percussive instrument), a *Transient Beat* sequence (synthesized music), and a polyphonic *Jazz* music snippet. All recordings are of studio quality (demo at [21]). The duration of files ranged between 4.5 s (for female speech) to 11.2 s (for *flute*), and all were digitized at $F_s = 16$ kHz. Sennheiser HD 215 headphones, which have almost flat frequency response between $0.02 - 8$ kHz, were used for listening. The listeners are first trained with a few examples of time-scaled signals to anchor their perception on quality rating. In the test set-up, four scaling factors $\alpha = \{0.5, 0.8, 1.2, 1.5\}$ are used. The listening test is carried with a GUI. The GUI presents for each scaling factor, and audio signal kind, the original, and five test signals coming from the five methods (random to the listener) for TSM simultaneously. The listener is allowed to play these signals any number of times, and is required to compare the TSM quality with the original signal as reference. A slider with an analog scale of $0-100$, and equi-partitioned by labels: *Bad, Poor, Good, Fair,* and *Best* is provided for rating each test signal. In total, each listener listened to $5 \times 7 \times 4 = 140$ test TSM files, and 7 original reference files. Ten listeners, in the age group of $21 - 35$ (all are university students and three being the co-authors of this paper), participated in the listening test. A listener on average took 45 mins for completing the test.

The mean rating (score), and std. deviation for the five methods after pooling data from all ten listeners is shown in Fig. 6(a). The μTVS has a mean rating close to 70, and together with the std. deviation, it is rated between *Fair-Best*. These ratings are better than those assigned to the other four methods. The μTVS method is preferred over the quasi-statinary OLA approaches, that is WOLA, PV, and PV-IP by considerable margin. The HPS method, being a strategic combination of WSOLA and PV-IP, outperforms both of them. However, it is rated slightly below the μTVS approach.

Interesting observations can be drawn by visualizing the ratings assigned for each time-scaling factor, and for different signals. The mean, and std. deviation of the ratings computed with 10 listeners for each - method, signal, and, α is shown in Fig. 6(b). The PV has a mean rating between *Bad-Poor* at all α, and for all signals except *Flute*, (though the std. deviation is high for it as well). Flute signal was found to be an exceptional signal in terms of being rated between *Fair-Good* at all four α, and for all methods. This is due to slow spectro-temporal modulations in *Flute* signal, and time-scaling such signals was found robust to TSM artifacts. The PV-IP method has better rating over PV, and this corroborates [5], highlighting the improved quality due to reduced phasiness perception. PV, and PV-IP both got minimal rating for *Glockenspiel*, and *Transient Beats*, at all α. This is because of smearing of sharp attacks by these methods. The WSOLA method performs quite well for $\alpha < 1$ but received a significant drop in rating for $\alpha > 1$. This is due to duplication of transient regions which results in annoying stuttering artifact in time-scaled signal. Further, for $\alpha > 1$ listeners slightly preferred smearing artifact over transient duplication, as can be observed by the improved rating of PV-IP over WSOLA. HPS method uses WSOLA for time-scaling of percussion stream but with a smaller OLA window size (16 ms) than that of WSOLA (64 ms). This results in reduced artifact of transient duplication for HPS, as an outcome, HPS received a better rating than WSOLA, and PV-IP. The μTVS approach has a more consistent performance across $\alpha < 1$ or > 1, the mean rating is better, and the std. deviation is low for most kinds of signals. The performance is significantly better for *Speech*, and *Vocals* signals. This is due to significantly less phasiness artifact, and no transient duplication. The performance of μTVS for *Transient Beats* is rated between *Fair-Good*, but falls short of HPS. This is due to some transient smearing artifact in the μTVS. Interestingly, the TSM on *Glockenspiel* is rated better with μTVS than HPS. This is likely because *Glockenspiel* has a sharp attack followed by a decay of the harmonic envelope. The little smearing of the attack portion introduced by μTVS is masked by the good quality TSM of the harmonic envelope decay. Features like these contribute towards higher rating of μTVS for polyphonic *Jazz* signal, as well.

IV. Conclusion

The results demonstrate improved performace in TSM obatined with mel-scale sub-band based time-varying sinusoidal model. Suprisingly, it outperforms the sophisticated quasi-stationary TSM methods. There is no requirement for windowing (and hence, no overhead of choice of window duration), and the approach is applicable to diverse kind of audio signals. Altering the IA, and IP via time-scaling does alter the temporal modulations embedded in them. However, based on the results of listening tests, we hypothesize that these alterations are perceived natural to the signals within the time-scaling factor of $0.5 - 1.5$. As direction for future

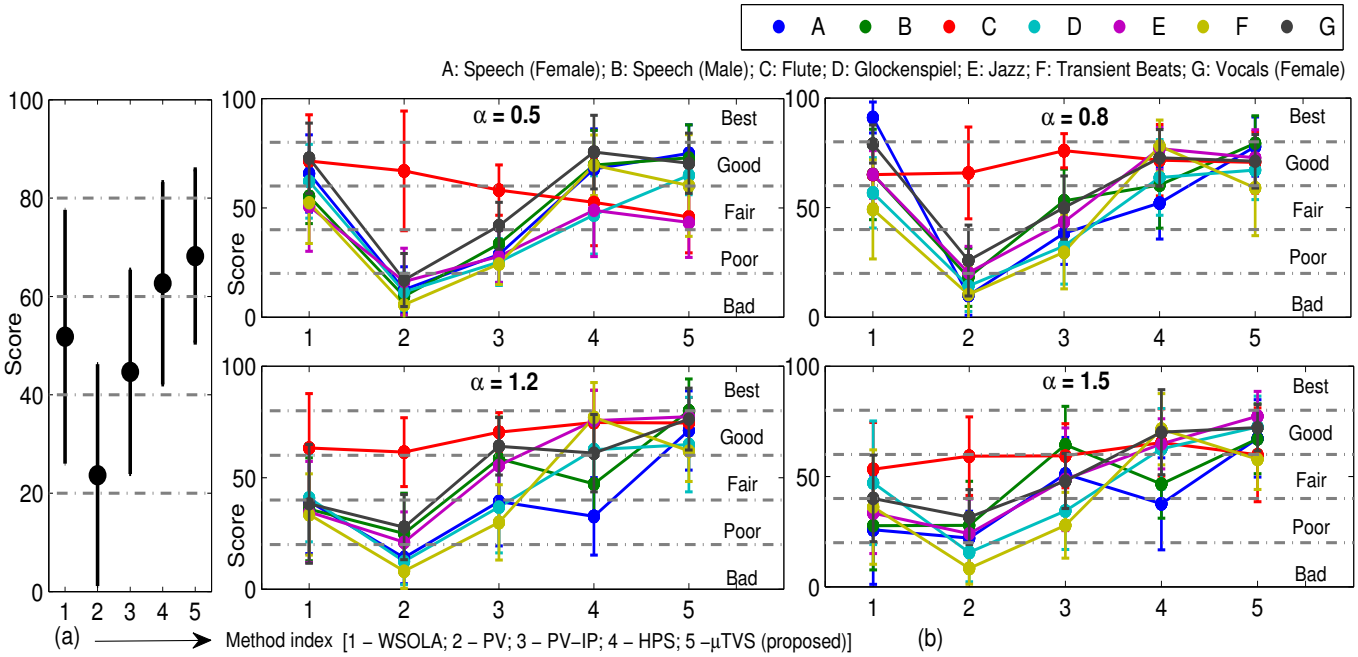

Figure 6: [In color] (a) Mean rating (score) by pooling the listening test results for each TSM method. (b) Signal-wise, and scaling factor-wise mean score obtained over all listeners. The data corresponds to ten listeners, and the error bars correspond to 1 std. deviation.

work, we note that the preservation of temporal structure of transients is not perfect. Also, the bandwidth of the filters increases progressively with filter index. Owing to this the narrowband assumption on sub-band signals will degrade over filter indices. We have found that μTSM technique does not preserve the time-frequency modulations for higher indices filters very well. Interestingly, this is not perceived as distortion, and needs more investigation.

References

[1] S. L. Hanauer and M. R. Schroeder, "Nonlinear time compression and time normalization of speech," *J. Acous. Soc. Amer.*, vol. 40, no. 5, pp. 1243–1243, 1966.

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

[3] 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 Intl. Conf. Acoust. Speech Signal Process.*, vol. 2, April 1993, pp. 554–557.

[4] M. R. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 29, no. 3, pp. 374–390, Jun 1981.

[5] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 323–332, May 1999.

[6] 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, Jan 2008.

[7] F. Nagel and A. Walther, "A novel transient handling scheme for time stretching algorithms," in *Audio Engineering Society Convention 127*, Oct 2009.

[8] J. Driedger, M. Muller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 105–109, Jan 2014.

[9] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010, pp. 1–4.

[10] H. Dudley, "Remaking speech," *J. Acous. Soc. Amer.*, vol. 11, no. 2, pp. 169–177, 1939.

[11] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, 1966.

[12] D. Vakman, "On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 791–797, Apr 1996.

[13] D. Gabor, "Theory of communication," *IEE J. Comm. Eng.*, vol. 93, pp. 429–457, 1946.

[14] N. K. Sharma and T. V. Sreenivas, "Event-triggered sampling using signal extrema for instantaneous amplitude and instantaneous frequency estimation," *Signal Processing*, vol. 116, no. C, pp. 43 – 54, 2015.

[15] P. Flandrin, "Time-frequency and chirps," *Proc. SPIE 4391, Wavelet Applications VIII*, pp. 161–175, Mar. 2001.

[16] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*. Berlin: Springer Verlag, Academic Press, 1990.

[17] T. Quatieri, R. Dunn, and T. Hanna, "A subband approach to time-scale expansion of complex acoustic signals," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 515–519, Nov 1995.

[18] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acous. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.

[19] L. Cohen and C. Lee, "Standard deviation of instantaneous frequency," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, vol. 4, May 1989, pp. 2238–2241.

[20] J. Driedger and M. Müller, "TSM Toolbox: MATLAB implementations of time-scale modification algorithms," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Erlangen, Germany, 2014, pp. 249–256.

[21] "A demo of μTVS approach for time-scale modifications," http://www.sagiisc.in/demos, accessed: 2016-08-05.