Proceedings of the Institute of Acoustics

USES OF THE PITCH-SCALED HARMONIC FILTER IN SPEECH PROCESSING

Philip J.B. Jackson School of Electronic and Electrical Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [p.jackson@bham.ac.uk]

Christine H. Shadle Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK. [chs@ecs.soton.ac.uk]

Abstract

The pitch-scaled harmonic filter (PSHF) is a technique for decomposing speech signals into their periodic and aperiodic constituents, during periods of phonation. In this paper, the use of the PSHF for speech analysis and processing tasks is described. The periodic component can be used as an estimate of the part attributable to voicing, and the aperiodic component can act as an estimate of that attributable to turbulence noise, i.e., from fricative, aspiration and plosive sources. Here we present the algorithm for separating the periodic and aperiodic components from the pitch-scaled Fourier transform of a short section of speech, and show how to derive signals suitable for time-series analysis and for spectral analysis. These components can then be processed in a manner appropriate to their source type, for instance, extracting zeros as well as poles from the aperiodic spectral envelope. A summary of tests on synthetic speech-like signals demonstrates the robustness of the PSHF's performance to perturbations from additive noise, jitter and shimmer. Examples are given of speech analysed in various ways: power spectrum, short-time power and short-time harmonics-to-noise ratio, linear prediction and mel-frequency cepstral coefficients. Besides being valuable for speech production and perception studies, the latter two analyses show potential for incorporation into speech coding and speech recognition systems. Further uses of the PSHF are revealing normally-obscured acoustic features, exploring interactions of turbulence-noise sources with voicing, and pre-processing speech to enhance subsequent operations.

1 Introduction

Voiced consonants and hoarse or breathy vowels contain significant contributions from both voicing and noise sources. Separating the speech signal into periodic and aperiodic components allows more accurate characterisation of each acoustic source [9] (for production models and articulatory synthesis). Separation also allows modification of the periodic component, as needed in concatenative synthesis, source-specific enhancement, e.g. [3], and diagnosis of pathologies through alternative representation of the noise. This article focuses on the application of the pitch-scaled harmonic filter (PSHF) as a precursor to a range of conventional processing techniques.

2 Method

Our decomposition technique, the PSHF, is based on a measure of harmonics-to-noise ratio derived by Muta et al. [8]. Calculating the HNR from a short section of speech $s(n)$, they used the spectral properties of an analysis frame scaled to the pitch period for distinguishing parts of the spectrum containing periodic energy from those without. Using a four pitch-period Hann window, $w(n) = 0.5\,(1 - \cos 2\pi n/N)$ for $n \in \{0, 1, \ldots, (N-1)\}$, centred at time p, they windowed $s(n)$ to form $s_w(n) = w(n)\,s(n+p-N/2)$, where the length $N(p) = bT_0(p)$ was a whole number of pitch periods T_0 (in samples, $b = 4$). The spectrum $S_w(k)$ was calculated by DFT, which concentrated the periodic part of s_w into the set of harmonic bins, $\mathcal{H} \in \{b, 2b, 3b, \ldots, b(N-1)\}$. Here, the harmonics are translated to bins $\{4, 8, 12, \ldots\}$. The value $b = 4$ was used because four is the smallest number that leaves bins free of spectral leakage from the periodic component (those half-way between the harmonics: $\{2, 6, 10, \ldots\}$), as shown in Figure 1. A larger value of b would make the decomposition more susceptible to degradation from the many kinds of variation in speech, e.g., in amplitude, fundamental frequency, formant frequencies, at voice onset/offset).[1] So, for speech with a pitch period of 6 ms and $b = 4$, the window would be 24 ms long.

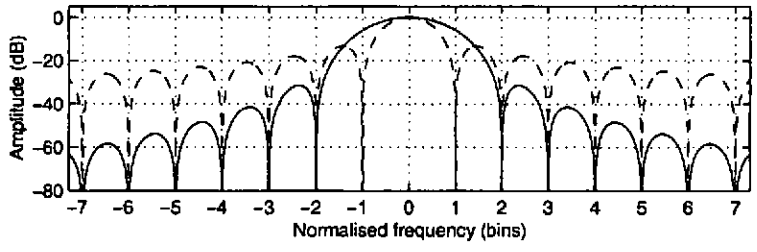

Figure 1: Comparison of the smearing effects on the spectral envelope of rectangular (solid) and Hann (dashed) windows.

The gain in robustness from using a Hann window, compared to a rectangular window, can be described by the sensitivity of cross-term bias errors between harmonics to deviations from perfect periodicity. These errors are reduced by a factor of 15 by the Hann window at the adjacent harmonic, four bins away. Also, the half-power bandwidth of the main peak at each harmonic is increased by 60 %, though it is related to an increase in estimation variance. In summary, a small part of the maximum likelihood performance for perfectly periodic signals is compromised to make the PSHF more suitable for time-varying signals.

While windowing allows the piecewise-stationary PSHF to adapt, we need to recombine the output signals after decomposition, which is achieved by overlapping and adding. However, we also need to normalise the aggregate back to unity gain, using the factor

$$W(p) = \frac{1}{\sum_i \left[w_i\,(p - p_i + N(p_i)/2)\right]}\,, \tag{1}$$

[1] In theory, any positive integer would suffice, though we have not experimented with any alternatives ourselves.

where the summation includes all windows w_i, centred at p_i, that contain time p. A cosine ramp was applied to each end of the normalisation factor $W(n)$ to fade out sections of voicing at onset and offset.

We have extended the process to yield a full decomposition into periodic (estimate of voiced) and aperiodic (estimate of unvoiced) complex spectra, which can be converted back into time series $\hat{v}$ and $\hat{u}$ respectively by inverse Fourier transformation and windowing, as explained below. We have also proposed an interpolation step for improving power-spectral estimation, which produces $\tilde{v}$ and $\tilde{u}$ [5, 7]. The signals can later be analysed using any standard technique (as will be demonstrated later in this paper): $\hat{v}$ and $\hat{u}$ for time-domain analysis, $\tilde{v}$ and $\tilde{u}$ for frequency-domain analysis. For time-frequency analysis, we define a threshold frame size of half the mean PSHF window length, $\langle N \rangle/2$ or two pitch periods, which is the point at which the harmonics begin to be resolved. Thus, $\hat{v}$ and $\hat{u}$ would be used for wide-band spectrograms, and $\tilde{v}$ and $\tilde{u}$ for narrow-band ones.

2.1 Pitch estimation

Requiring that the window length N be scaled to the time-varying pitch period $T_0(p)$ means that we must first estimate the pitch period of any speech signal. Initial estimates may be obtained manually or from the autocorrelation, for instance. The pitch-estimation algorithm then finds optimum value, for a particular time, by minimising the amount of smearing in the spectrum from the first H harmonics, $h \in \{1, 2, \ldots, H\}$. Thus, the spectral sharpness is expressed in terms of the higher and lower spectral spread, S_h^+ and S_h^- respectively, as defined in [8], and the cost function is

$$J(N,p) = \sum_{h=1}^{H} \left[S_h^+(N,p)^2 + S_h^-(N,p)^2 \right], \tag{2}$$

providing a pitch estimate perfectly matched to the decomposition, as described below. See [8] for further details. Thus, a pitch track can be obtained by repeating this process at successive points throughout the speech signal.

2.2 Algorithm: Harmonic filter

Let us consider how the PSHF decomposes a single frame, centred at time p, in its harmonic filter, the central engine of the algorithm. After applying the pitch-scaled Hann window to the speech signal, $s_w(n)$ is discrete Fourier transformed (DFT) to $S_w(k)$, as depicted by $\mathcal{F}$ in Figure 2.

The harmonic filter extracts the pitch harmonics from S_w, doubling their coefficients to compensate for the mean window amplitude, which form the harmonic spectrum:

$$\hat{V}(k) = \begin{cases} 2S_w(k) & \text{for } k \in \mathcal{H} \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathcal{H} = \{4, 8, \ldots, 4(N-1)\}$. Once returned to the time domain by inverse DFT (IDFT), these four

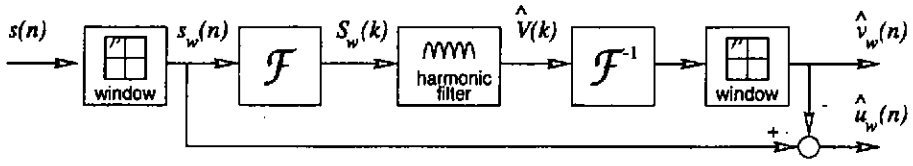

Figure 2: The pitch-scaled harmonic filter (PSHF) comprising Hann window, discrete Fourier transform ($\mathcal{F}$), harmonic filter, inverse DFT ($\mathcal{F}^{-1}$) and another Hann window operation to yield the outputs for time-series processing, $\hat{v}_w$ and $\hat{u}_w$.

pitch periods are windowed to yield the periodic signal estimate:

$$\hat{v}_w(n) = \frac{w(n)}{N} \sum_{k=0}^{N-1} \left[\hat{V}(k) \exp\left(j \frac{2\pi n k}{N} \right) \right] . \tag{4}$$

The aperiodic signal estimate is the difference between the input signal and the periodic estimate: $\hat{u}_w(n) = s_w(n) - \hat{v}_w(n)$. This could be obtained equivalently via the frequency domain: $\hat{U}(k) = S(k) - \hat{V}(k)$, by IDFT and windowing.[2] As a result of the subtraction, any errors in the periodic estimate caused by the decomposition algorithm are (wrongly) attributed to the aperiodic signal.

2.3 Algorithm: Power interpolation

The spectrum of the (windowed) aperiodic signal, $\hat{U}_w(k)$, contains gaps at the pitch harmonics where the coefficients are of zero amplitude, since $\hat{U}_w(k) = S_w(k) - \hat{V}_w(k) = S_w(k) - (2S_w(k))/2 = 0$ for $k \in \mathcal{H}$.[3] However, subsequent analysis often involves spectral magnitudes, power spectra or spectrograms, and so the gaps would give strongly biased under-estimates. We can improve these estimates by patching suitable values into $\hat{U}_w$ at the harmonics, according to the process shown in Figure 3.

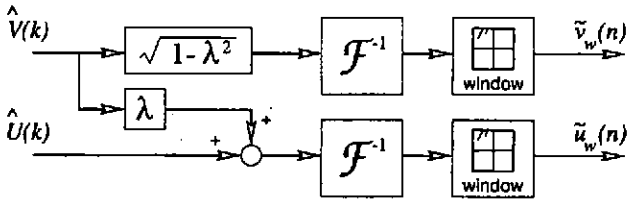

Figure 3: The PSHF's interpolation stage, which redistributes the power at the pitch harmonics to produce the outputs for spectral processing, the components $\tilde{v}_w$ and $\tilde{u}_w$.

Assuming the aperiodic component to be stochastic with a smooth frequency response, the power in any frequency bin k is expected to be similar to that of the adjacent bins $k \pm 1$. Hence, we calculate

[2] Unwindowed spectra, $S(k)$, $\hat{V}(k)$ and $\hat{U}(k)$, correspond to unwindowed signals in the time domain: $s(n)$, $\hat{v}(n)$ and $\hat{u}(n)$.
[3] $\hat{V}_w(k)$ is half of $\hat{V}(k)$ at the harmonics owing to the smearing effect of the Hann window.

$L(k)$, a frequency-local estimate of $|U_w(k)|$ at the harmonics, by interpolation of the adjacent powers:

$$L(k) = \sqrt{\frac{1}{2}\left(\left|\hat{U}_w(k-1)\right|^2 + \left|\hat{U}_w(k+1)\right|^2\right)} \qquad \text{for } k \in \mathcal{H}. \tag{5}$$

The factor $\lambda(k)$ determines the distribution of $S_w(k)$'s power between the revised periodic and aperiodic estimates by comparing it with $L(k)$'s power, thus:

$$\lambda(k) = \frac{L(k)}{\sqrt{|S_w(k)|^2 + (L(k))^2}}. \tag{6}$$

The revised estimates are:

$$\tilde{V}(k) = \begin{cases} \sqrt{1-\lambda(k)^2}\,\hat{V}(k) & \text{for } k \in \mathcal{H}; \\ \hat{V}(k) & \text{otherwise,} \end{cases} \quad \text{and} \quad \tilde{U}(k) = \begin{cases} \hat{U}(k) + \lambda(k)\hat{V}(k) & \text{for } k \in \mathcal{H}; \\ \hat{U}(k) & \text{otherwise.} \end{cases} \tag{7}$$

Note that using the original phase information for both components, $\arg(S_w(k))$, enables us to reconstruct the power-based time series $\tilde{v}_w(n)$ and $\tilde{u}_w(n)$ (by IDFT and windowing, as before), so that consistency is maintained across overlapping frames. These signals retain the detail of the original time series, while avoiding artefactual troughs at pitch harmonics in the magnitude spectrum. Each of the four outputs, $\hat{v}_w$, $\hat{u}_w$, $\tilde{v}_w$ and $\tilde{u}_w$, can be overlapped with outputs from neighbouring frames to yield two pairs of contiguous periodic and aperiodic components: $\hat{v}$ and $\hat{u}$, and $\tilde{v}$ and $\tilde{u}$.

3 Evaluation

The performance of the PSHF was evaluated using speech-like test signals $s(n)$, which were made up of a synthetic voiced part $v(n)$ and unvoiced part $u(n)$: $s(n) = v(n) + u(n)$. The sampling rate was $f_s = 48$ kHz. The voiced part was produced by convolving a periodic pulse train $g(n)$ with an appropriate impulse-response filter $q(n)$: $v = g * q$. The filter q was built using the linear prediction coefficients (LPC, 50-pole, autocorrelation) obtained from a recorded adult male [aa] vowel. For jitter and shimmer tests, the pulses were randomly perturbed from their nominal amplitude and pitch period by amounts ranging from 0 to 5 % and 0 to 1.5 dB, respectively [4]. The unvoiced part was similarly created by convolving Gaussian white noise $d(n)$ (zero mean, unit variance) with the LPC filter: $u = A\,d * q$, and the gain A was adjusted to give the desired HNR (-5 to ∞ dB). Tests using modulated noise are reported elsewhere [5, 6].

So, using the PSHF component estimates $\hat{v}$ and $\hat{u}$, the changes in signal-to-error ratio (SER) were calculated, as a measure of the decomposition algorithm's performance. The change in SER for the periodic component η_p is defined as the ratio of the unvoiced part's mean power to that of the residual error, $e = (\hat{v} - v) = -(\hat{u} - u)$; conversely, the aperiodic performance η_a is the ratio of voiced to residual-error power (expressed in dB):

$$\eta_p = 10\log_{10}\left[\langle u^2\rangle / \langle e^2\rangle\right], \qquad \text{and} \qquad \eta_a = 10\log_{10}\left[\langle v^2\rangle / \langle e^2\rangle\right]. \tag{8}$$

Similarly, we define an estimate of the HNR as: $\rho = 10\log_{10}\left[\langle \tilde{v}^2\rangle / \langle \tilde{u}^2\rangle\right]$.

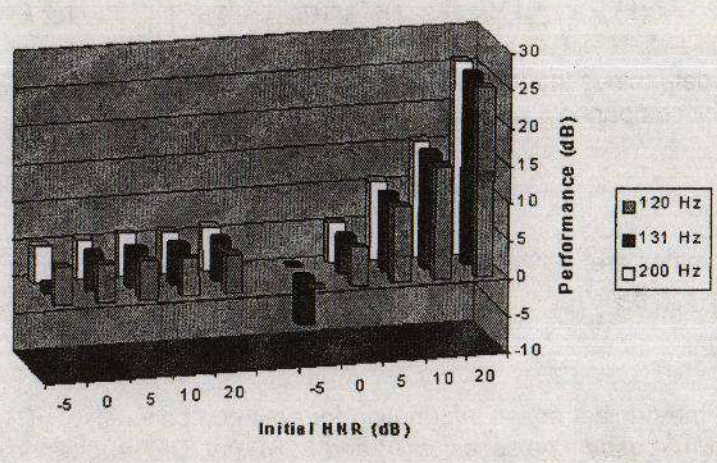

Figure 4: Performance of PSHF for the periodic (left) and aperiodic (right) components, η_p and η_a respectively (with no jitter or shimmer). Results are shown here for three values of f_0 (120.0 Hz, 130.8 Hz, 200.0 Hz), and five values of initial HNR (−5 dB, 0 dB, 5 dB, 10 dB and 20 dB).

3.1 Test results

Figure 4 shows the effects of the initial HNR and fundamental frequency on performance. For all but one extreme exception (−5 dB, $f_0 = 130$ Hz), the fundamental frequency has a negligible influence, because the resolution of the harmonic filter is effectively constant relative to f_0. For the periodic component, the effect of HNR is marginal, though it degrades slightly under very noisy conditions, while the aperiodic performance is very dependent on the initial HNR. The results show that the PSHF enhanced the estimate of the voiced component by 5 to 6 dB, and the aperiodic estimate by approximately 5 dB more than the HNR.

With frequency and amplitude perturbations from jitter and shimmer respectively included, there was a similar pattern, albeit increasingly degraded as the degree of perturbation was increased. The results in Figure 5 show that, for perturbation levels normally found in speech, the performances are typically reduced by approximately 1 dB, compare to the unperturbed results. Yet, improvements to the estimate of the voiced part were obtained with severe jitter, shimmer and noise, e.g. (J, S, HNR): (3 %, 1 dB, 5 dB) and (3 %, 0 dB, 10 dB). Fluctuations in the pitch period (jitter) appeared to have a larger effect on the PSHF's performance than amplitude fluctuations (shimmer).

A list of performance results and estimated HNRs is given in Table 1 for completeness. At very high HNRs ($\rightarrow \infty$ dB), minor changes in the random noise and quantisation can have an exaggerated effect on the variance of error measurements.

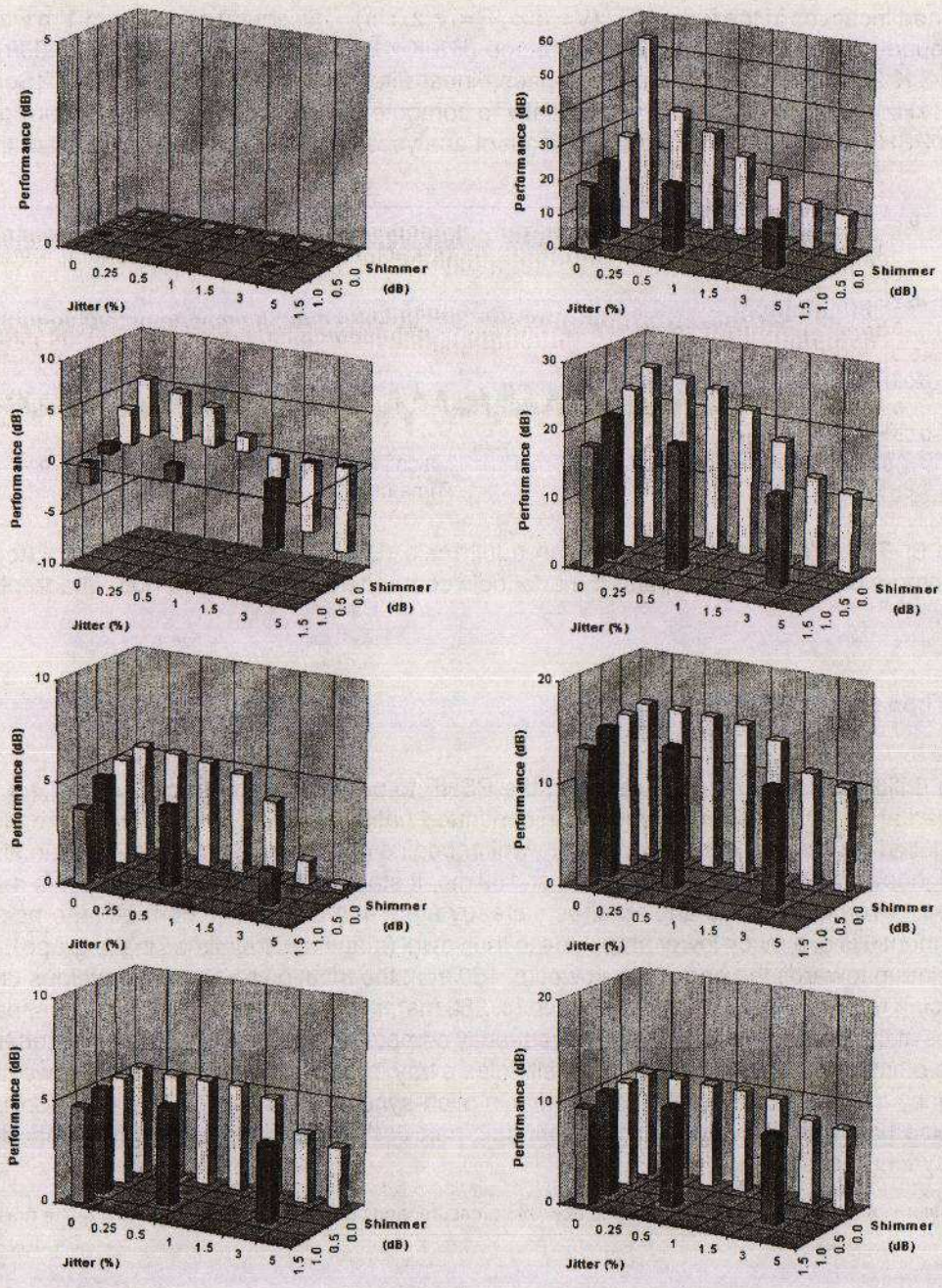

Figure 5: Performance of PSHF for the periodic (left) and aperiodic (right) components, η_p and η_a respectively, versus jitter and shimmer, for differing HNRs: (from top) ∞ dB, 20 dB, 10 dB and 5 dB. Note that the vertical scale varies. ($f_0 = 130.8$ Hz.)

4 Secondary analyses

An adult male (LJ), a native speaker of European Portuguese, recorded a speech corpus that included sustained fricatives in the form /VF:/ (V = /ax/, F = /v, z, zh/). The sound pressure at 1 m was measured in a sound-treated booth using a microphone (B & K 4165), a pre-amplifier (B & K 2639) and amplifier (B & K 2636, 22 Hz to 22 kHz band-pass, linear filter). It was recorded on DAT (Sony TCD-D7, $f_s = 48$ kHz), and thence transferred digitally to computer.[4] These speech signals were decomposed by the PSHF in preparation for the subsequent analyses described in the remainder of this section.

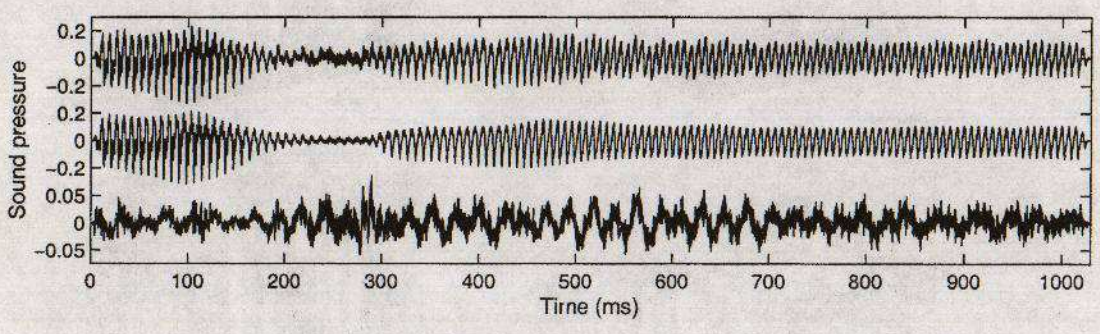

Figure 6: Time series from [ax-v:] by an adult male (LJ) of the original signal $s(n)$ (top), the periodic component $\hat{v}(n)$ (middle), and the aperiodic component $\hat{v}(n)$ (bottom, note quadruple amplitude scale).

4.1 Time series

Figure 6 illustrates the result of applying the PSHF to part of the utterance [ax-v:]. The voice onset occurred at $t = -100$ ms and the fricative continued until $t = 5.1$ s. The majority of the signal energy is modelled by the periodic component $\hat{v}$, which begins mid-vowel during voicing, and is maintained at a high level throughout the vowel. After 100 ms, it starts to fade as the transition is made into the fricative, which overshoots and reaches a steady state at c. 320 ms. In contrast, the noisy aperiodic component $\hat{u}$ is of a much lower amplitude in the vowel (magnified four times in the graph), decaying to its minimum towards the end of the vowel (c. 150 ms); the frication noise then develops up to 260 ms. As voicing returns in the sustained fricative (c. 280 ms), there are some transient errors manifested as glitches at the pitch epochs. A very low-frequency component is perhaps evidence of flutter (combined effects of jitter and shimmer) that gradually dies away, particularly after 700 ms. However, the high-frequency frication noise continues to arrive in pitch-synchronous pulses. The waveforms show how the noise has been removed from the periodic part $\tilde{v}$; the noise appears in the aperiodic part $\tilde{u}$ as pitch-synchronous packets.

[4] A calibration tone was recorded to refer to absolute pressure, and background noise for assessing the noise floor.

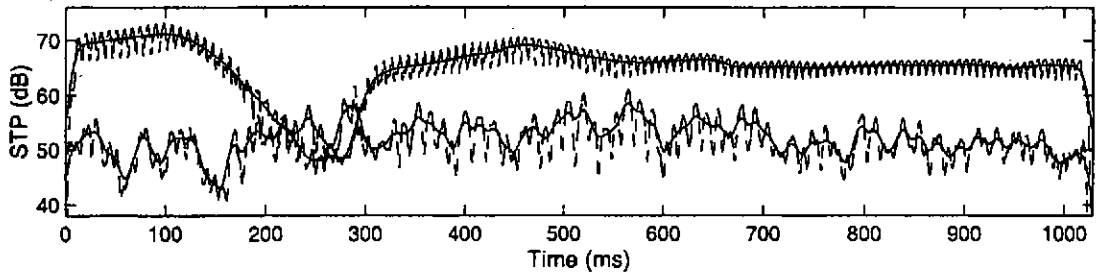

Figure 7: Short-time powers of the periodic (upper) and aperiodic (lower) components, P_p and P_a, during [ax-v:] by LJ. The dashed lines show short-term variations within each pitch period ($M \approx 7$ ms, Hann window); the solid lines medium-term ($M \approx 30$ ms, Hann window).

4.2 Short-time power

The short-time power (STP) is a quantity derived by calculating a moving, weighted average of the squared signal. It is defined, for any signal $y(n)$, as:

$$P_y(n) = \frac{\sum_{m=0}^{M-1} [x(m)^2 y(n+m-M/2)^2]}{\sum_{m=0}^{M-1} [x(m)^2]}, \tag{9}$$

using a smoothing window $x(m)$, which was set to a fixed length. The STP can be used to examine the pulsing of the noise component, as in [6], but if one is interested in medium-term effects (i.e., at a similar rate to articulatory movements) the oscillation can be removed by averaging over a few pitch periods. This point is illustrated in Figure 7 where the short-term STP ($M = \langle N \rangle$, where $\langle\,\rangle$ denotes the time-average) and medium-term STP ($M = 4\langle N \rangle$) are plotted. The STP of the periodic component, P_p, and that of the aperiodic component, P_a, were calculated using $\hat{v}$ and $\hat{u}$ respectively. While confirming our earlier observations, these trajectories provide a clear picture of the timing and relative amplitudes of the different acoustic contributions. Even for this weak fricative, the aperiodic STP shows a doubling of the noise amplitude on average following the vowel-fricative transition. The difference between the periodic and aperiodic STP curves gives a measure of the short-time HNR, which could be used as an objective means to determine the start of the fricative (e.g., for coarticulation studies).

4.3 Power spectra with linear predictive smoothing

Power spectra were calculated from a steady section of the speech signal and the power-based signals (s, $\tilde{v}$ and $\tilde{u}$, centred on 900 ms in Figs. 6/7), and are plotted in Figure 8 (upper half). Most of the energy in the original spectrum comes in the first four harmonics (of $f_0 \approx 140$ Hz) but there is a significant proportion above 3 kHz, even though the spectrum becomes more noisy at higher frequencies. However, the periodic spectrum maintains its harmonic structure over all frequencies shown, while

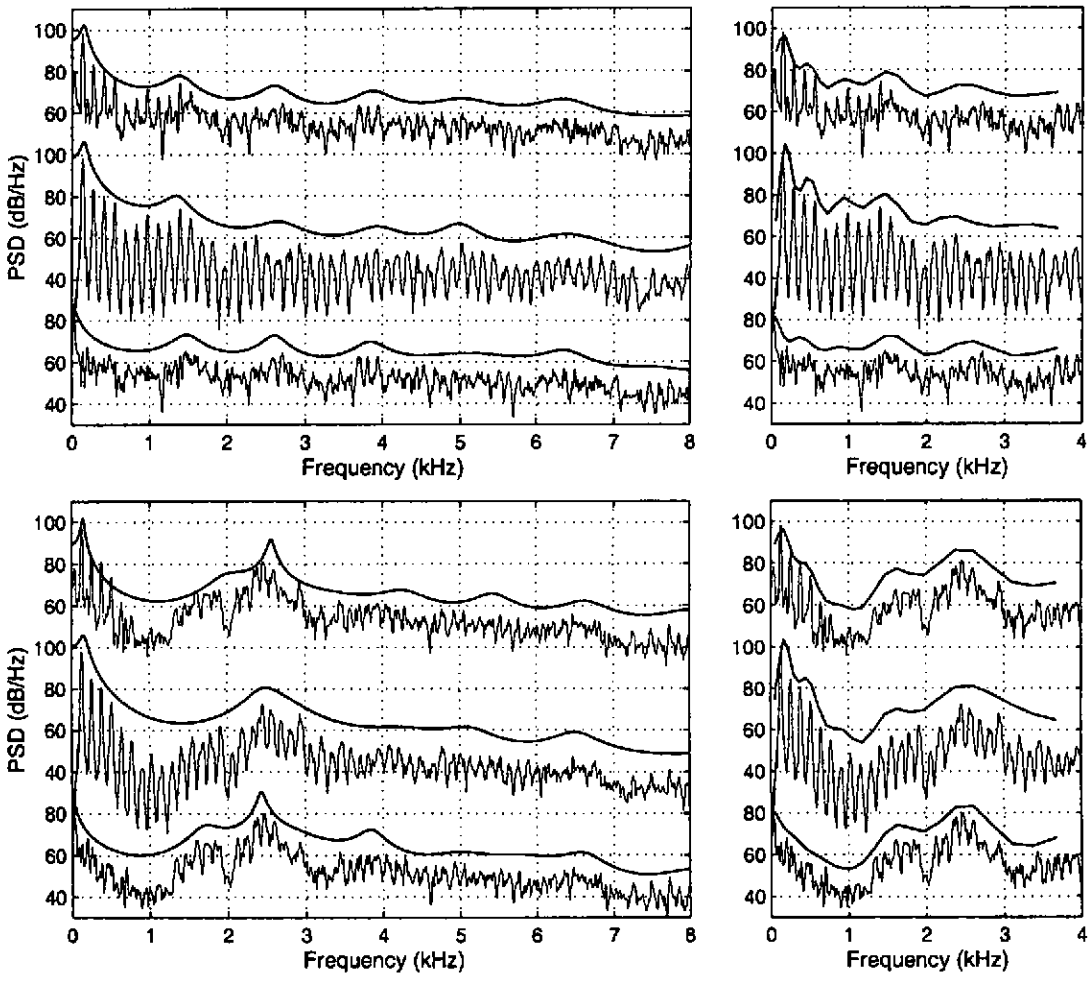

Figure 8: Power spectral density (85 ms, Hann window centred at 900 ms, $4\times$ zero-padded, averaged over 170 ms) computed mid-phone for the sustained fricatives by an adult male subject (LJ), (upper) [v:] and (lower) [zh:], using: (top) the original signal $s(n)$, (middle) the periodic estimate $\tilde{v}(n)$ and (bottom) the aperiodic estimate $\tilde{u}(n)$. The smooth overlaid approximations of the spectral envelope (thick lines) were calculated using (left) linear prediction coefficients and (right) Mel-frequency cepstral coefficients.

the aperiodic spectrum, being pervasively noisy, is almost completely devoid of harmonics. On the left, smoothed spectra derived from the linear prediction coefficients (50-pole, autocorrelation [1]) are superimposed. They demonstrate how both components differ from the original, in terms of spectral tilt and the frequency and bandwidth of the spectral peaks, e.g., in kHz: original (0.15, 1.38, 2.61), periodic (0.15, 1.34, 2.65), aperiodic (absent, 1.48, 2.61). For comparison, we refer to the peaks as F0, F2 and F3, respectively, and the less-prominent hump near the third harmonic (c. 420 Hz) as F1. Note that the periodic and aperiodic LPC spectra first transect just above 2 kHz because of their tilts, a lower frequency than might be expected.

Figure 8 (lower half) shows the fricative [zh:] uttered by the same subject, having been processed similarly. Its LPC spectra (left) transect lower still at c. 1.3 kHz as the frication noise dominates from F2 upwards (see also [z:] by PJ [7]). Peak frequencies (F0, F2, F3) in kHz are: original (0.14, (2.02), 2.57), periodic (0.15, -, 2.49), aperiodic (-, 1.74, 2.44). The bandwidths also differ considerably (cf. F3), and formants are both inserted and deleted with respect to the original.

4.4 Mel-frequency cepstral smoothing

Mel-frequency cepstral coefficients (MFCCs) were calculated according to standard procedures [10]. The binning process used an array of band-pass filters that were triangular and equally-spaced between 0 and 4 kHz on the Mel-frequency scale:

$$f_{\text{Mel}} = 1127 \ln\left(1 + f_{\text{Hz}}/700\right) , \tag{10}$$

where ln implies the natural logarithm. This yields a value for the Mel-frequency log-magnitude spectrum, $S(\kappa)$, at each Mel-frequency bin κ.

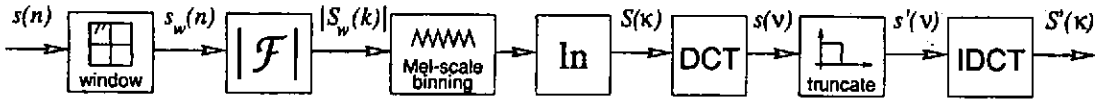

Figure 9: Calculation of smoothed log spectra via Mel-frequency cepstral coefficients (MFCCs).

As shown in Figure 9, the discrete cosine transform (DCT) is applied to $S(\kappa)$, to form the MFCCs, $s(\nu)$ where ν is the coefficient index [1]. These are truncated, as denoted by prime $s'(\nu)$, keeping only thirteen (13) values and setting the remainder equal to zero (the zeroth coefficient is also often discarded), and finally the inverse DCT (IDCT) applied to yield $S'(\kappa)$. The frequency-warping (and non-linear frequency weighting) has been removed for the plots in Figure 8 (right). In summary,

$$s(\nu) = c(\nu)\sqrt{\frac{2}{K}}\sum_{\kappa=0}^{K-1}\left[S(\kappa)\cos\left(\frac{(2\kappa+1)\,\pi\nu}{2K}\right)\right] ; \tag{11}$$

$$S'(\kappa) = C(\kappa)\sqrt{\frac{2}{K}}\left[\sum_{\nu=0}^{K-1} s'(\nu)\cos\left(\frac{(2\nu+1)\,\pi\kappa}{2K}\right)\right] ; \tag{12}$$

where $c(\nu) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } \nu = 0; \\ 1 & \text{otherwise,} \end{cases}$ and $C(\kappa) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } \kappa = 0; \\ 1 & \text{otherwise.} \end{cases}$

The smoothed spectra on the right-hand side of Figure 8 are derived from Mel-frequency cepstral coefficients (MFCCs). Here, the amount of smoothing increases with frequency, but still the three curves are distinct for each example. The original MFCC spectrum has separate peaks near the first harmonic (F0 $\approx$ 160 Hz) and the next peak above that (F1 $\approx$ 450 Hz), as does the periodic MFCC spectrum. However, F3 is lower for the periodic component, compared with the original. The aperiodic MFCC spectrum also has a peak at F1 (but not at F0), though rather low in frequency (c. 370 Hz), and again its F2 is higher than the others.

For [zh:], the differences between the aperiodic and the other two MFCC spectra are yet more pronounced at low frequencies (<500 Hz). The MFCC spectra generally fit better than the LPC ones (again more apparent for [zh:]), especially in the vicinity of spectral troughs. These disparities suggest that more radical automatic speech recognition systems that can use different models for different classes of acoustic source (e.g., [2]) may be able to gain a significant advantage by processing these components in parallel.

5 Conclusion

We have presented a signal decomposition technique, its evaluation using synthetic signals, and results from its application to real speech. The PSHF enables separate analyses of periodic and aperiodic components as estimates of the voiced and unvoiced parts in mixed-source speech. The best decompositions are achieved during sustained phonation, since the PSHF is based on a harmonic model. Although severe jitter and shimmer can induce artefacts in the aperiodic component, the evaluation indicated consistent improvements over typical conditions. Short-time power, LPC and MFCC analyses showed the potential for modelling of these components individually. Coupled with the examples given, they demonstrate that the PSHF algorithm offers novel forms of feature extraction for speech when turbulence noise and voicing co-occur.

References

[1] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-time Processing of Speech Signals.* Macmillan, NJ, 1993.

[2] L. Deng and J. Ma. A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics. 4:1499–1502, 1999.

[3] J. Hardwick, C. D. Yoo, and J. S. Lim. Speech enhancement using the dual excitation speech model. *Proc. IEEE-ICASSP*, 2:367–370, 1993.

[4] J. Hillenbrand. A methodological study of perturbation and additive noise in synthetically generated voice signals. *J. Speech & Hearing Res.*, 30:448–461, 1987.

[5] P. J. B. Jackson. *Characterisation of plosive, fricative and aspiration components in speech production.* PhD thesis, Dept. Electronics & Comp. Sci., Univ. of Southampton, Southampton, UK, 2000. http://web.bham.ac.uk/p.jackson/abstracts.html.

[6] P. J. B. Jackson and C. H. Shadle. Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. *J. Acoust. Soc. Am.*, 108(4):1421–1434, 2000.

[7] P. J. B. Jackson and C. H. Shadle. Performance of the pitch-scaled harmonic filter and applications in speech analysis. *Proc. IEEE-ICASSP*, Istanbul, 3:1311–1314, 2000.

[8] H. Muta, T. Baer, K. Wagatsuma, T. Muraoka, and H. Fukuda. A pitch-synchronous analysis of hoarseness in running speech. *J. Acoust. Soc. Am.*, 84(4):1292–1301, 1988.

[9] S. Narayanan and A. Alwan. Parametric hybrid source models for voiced and voiceless fricative consonants. *Proc. IEEE-ICASSP*, 1:377–380, 1996.

[10] S. J. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book.* Cambridge Univ. Tech. Services Ltd., Cambridge, UK, version 2.1 edition, 1995. http://htk.eng.cam.ac.uk/.

HNR dB	J %	S dB	f_0 Hz	η_p dB	η_a dB	ρ dB	HNR dB	J %	S dB	f_0 Hz	η_p dB	η_a dB	ρ dB
−5	0	0	120	5.05	0.02	−1.02	10	0	1.5	131	3.69	13.48	10.04
−5	0	0	200	4.91	−0.20	−2.25	10	0.5	1	131	3.90	13.80	10.28
−5	0	0	131	1.04	−6.48	−2.48	10	3	1	131	1.63	11.48	7.41
0	0	0	120	5.03	5.00	2.49	20	0	0	120	5.26	25.23	21.59
0	0	0	200	5.36	5.26	1.22	20	0	0	200	5.93	25.82	21.14
0	0	0	131	4.99	4.90	1.19	20	0	1	200	1.19	21.03	18.32
5	0	0	120	5.15	10.12	6.95	20	0	0	131	5.67	25.58	21.19
5	0	0	200	5.73	10.62	6.19	20	0.25	0	131	4.81	24.72	20.71
5	0	1	200	5.49	10.32	6.01	20	0.5	0	131	3.89	23.80	20.39
5	0	0	131	5.31	10.22	6.06	20	1	0	131	1.55	21.46	18.22
5	0.25	0	131	5.22	10.13	5.92	20	1.5	0	131	−2.22	17.68	15.12
5	0.5	0	131	5.20	10.11	6.15	20	3	0	131	−6.83	13.07	10.88
5	1	0	131	5.19	10.09	6.13	20	5	0	131	−8.22	11.68	8.82
5	1.5	0	131	4.86	9.77	5.73	20	0	0.5	131	3.55	23.46	19.77
5	3	0	131	3.38	8.28	5.16	20	0	1	131	0.92	20.82	18.01
5	5	0	131	2.97	7.88	4.31	20	0	1.5	131	−2.13	17.65	15.73
5	0	0.5	131	5.24	10.15	5.94	20	0.5	1	131	−1.61	18.29	16.12
5	0	1	131	5.12	10.02	5.86	20	3	1	131	−6.42	13.44	10.50
5	0	1.5	131	4.68	9.46	5.56	∞	0	0	120	−∞	72.70	72.70
5	0.5	1	131	4.79	9.69	5.84	∞	0	0	200	−∞	49.74	49.74
5	3	1	131	3.79	8.64	4.21	∞	0	1	200	−∞	22.77	21.50
10	0	0	120	5.15	15.12	11.74	∞	0	0	131	−∞	54.05	54.05
10	0	0	200	5.74	15.64	11.17	∞	0.25	0	131	−∞	34.04	30.91
10	0	1	200	5.00	14.84	10.74	∞	0.5	0	131	−∞	29.31	27.42
10	0	0	131	5.43	15.34	11.12	∞	1	0	131	−∞	23.70	21.41
10	0.25	0	131	5.31	15.22	10.99	∞	1.5	0	131	−∞	18.48	16.68
10	0.5	0	131	5.17	15.08	11.10	∞	3	0	131	−∞	13.39	11.25
10	1	0	131	4.84	14.75	10.77	∞	5	0	131	−∞	11.95	8.98
10	1.5	0	131	3.82	13.73	9.87	∞	0	0.5	131	−∞	27.55	26.40
10	3	0	131	1.16	11.06	8.22	∞	0	1	131	−∞	22.42	21.14
10	5	0	131	0.27	10.18	6.90	∞	0	1.5	131	−∞	18.48	17.23
10	0	0.5	131	5.19	15.10	10.87	∞	0.5	1	131	−∞	19.32	17.64
10	0	1	131	4.74	14.64	10.61	∞	3	1	131	−∞	13.68	11.28

Table 1: Periodic and aperiodic performance of the PSHF (η_p, η_a in dB) versus specified jitter (J), shimmer (S), fundamental frequency (f_0) and initial harmonics-to-noise ratio (HNR). Estimated HNRs ρ, derived from the outputs $\tilde{v}(n)$ and $\tilde{u}(n)$, are also given.