PERFORMANCE OF THE PITCH-SCALED HARMONIC FILTER AND APPLICATIONS IN SPEECH ANALYSIS

Philip J.B. Jackson and Christine H. Shadle

Dept. of Electronics and Computer Science, Univ. of Southampton, Southampton SO17 1BJ, UK.
Email: {pjbj96r, chs}@ecs.soton.ac.uk, Internet: http://www.isis.ecs.soton.ac.uk/

ABSTRACT

The pitch-scaled harmonic filter (PSHF) is a technique for decomposing speech signals into their voiced and unvoiced constituents. In this paper, we evaluate its ability to reconstruct the time series of the two components accurately using a variety of synthetic, speech-like signals, and discuss its performance. These results determine the degree of confidence that can be expected for real speech signals: typically, 5 dB improvement in the signal-to-noise ratio of the harmonic component and approximately 5 dB more than the initial harmonics-to-noise ratio (HNR) in the anharmonic component. A selection of the analysis opportunities that the decomposition offers is demonstrated on speech recordings, including dynamic HNR estimation and separate linear prediction analyses of the two components. These new capabilities provided by the PSHF can facilitate discovering previously hidden features and investigating interactions of unvoiced sources, such as frication, with voicing.

1. INTRODUCTION

There are many reasons for wanting to separate the voiced and unvoiced components of speech: to allow accurate characterization of each acoustic source [1] (for production models and articulatory synthesis); for modification, as used in concatenative synthesis; for enhancement, e.g. [2]; to aid diagnosis of pathologies through improved representation of the noise.

Our decomposition technique is based on a measure of HNR derived by Muta *et al.* [3]. They applied a pitch-scaled Hanning window, centered around time p, to the input speech signal: $s_w(n) = w(n)\, s(n+p-N/2)$, where $w(n) = 0.5\,(1-\cos 2\pi n/N)$ for $n \in \{0, 1, \ldots, N-1\}$, and the number of points, $N = b\,T_0$, is an integer multiple b of pitch periods T_0. In order to give good time resolution and to take advantage of the window's spectral properties, $b = 4$ was chosen. Then, the spectrum $S_w(k)$ was computed by discrete Fourier transformation (DFT) of the four whole pitch periods:

$$S_w(k) = \sum_{n=0}^{N-1} s_w(n) \exp\left(-j\frac{2\pi nk}{N}\right), \qquad (1)$$

which concentrated the periodic part of the signal into the harmonic coefficients, $\mathcal{H} = \{b, 2b, \ldots, b(N-1)\}$, every fourth DFT bin. Hence, the voiced component was modeled by an adaptive series of coefficients at the fundamental frequency f_0 and its harmonics, and the unvoiced component, assumed to be the product of a stochastic process, was the remainder. We have extended the process to yield a full decomposition of the speech signal into harmonic (voiced) and anharmonic (unvoiced) time series [4], to which standard analysis techniques can be applied subsequently. We also proposed an interpolation stage [5] for improving spectral estimation over longer time scales (i.e. finer frequency resolution, $\Delta f < f_0/2$), which is a crucial step.

2. METHOD

2.1. Pitch estimation

The PSHF requires that the window length N be scaled to the time-varying pitch period $T_0(p)$, which was estimated by sharpening the spectral peaks of the first H harmonics, according to the cost function:

$$J\,(N,p) = \sum_{h=1}^{H} S_h^+\,(N,p)^2 + S_h^-\,(N,p)^2, \qquad (2)$$

where S_h^+ and S_h^- respectively are the estimation errors for the higher and lower spectrum spread of the harmonics $h \in \{1, 2, \ldots, H\}$, for the given window length N (see [3] for details). For each section of voiced speech analyzed, the pitch-tracker was initialized manually and the optimum window size $N(p)$ was taken at the local minimum of the cost function. The window lets the algorithm kernel deal with small sections of the speech signal sequentially. Thus, the outputs can be combined

with previously processed ones to yield continuous harmonic and anharmonic signals.

2.2. The PSHF kernel

The kernel of the PSHF, as depicted in Figure 1, optimally separates the harmonic and anharmonic signal components in the frequency domain. So, having been

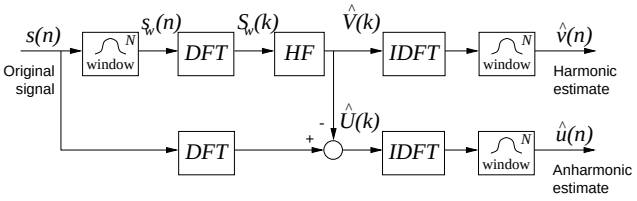

Figure 1: The pitch-scaled harmonic filter (PSHF).

windowed, the speech signal $s_w(n)$ undergoes a DFT to $S_w(k)$. In estimating the voiced spectrum $\hat{V}(k)$, the harmonic filter (HF) takes the complex amplitudes of the harmonics and doubles them (to counteract the mean window amplitude), which leaves the unvoiced estimate $\hat{U}(k)$:

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

$$\hat{U}(k) = \begin{cases} S(k) - 2S_w(k) & \text{for } k \in \mathcal{H}, \\ S(k) & \text{otherwise.} \end{cases} \quad (4)$$

The harmonic signal is calculated by inverse transforming (IDFT) and windowing, thus:

$$\hat{v}(n) = \frac{w(n)}{N} \sum_{k=0}^{N-1} \hat{V}(k) \exp\left(j \frac{2\pi n k}{N}\right), \quad (5)$$

and likewise for the anharmonic signal $\hat{u}(n)$. Now, although these signals are optimal in the time domain (in a least-squares sense), the anharmonic spectrum has gaps in it — at the harmonics — which may be misleading for frequency domain interpretations (power spectra, spectrograms, etc.). Therefore, for later analysis over long frame sizes ($> 2\,T_0$), the harmonic bins of the unvoiced spectrum are filled by interpolation. First, the (rms) average of the adjacent bins is calculated, for $k \in \mathcal{H}$:

$$L(k) = \sqrt{\frac{\left|\hat{U}_w(k-1)\right|^2 + \left|\hat{U}_w(k+1)\right|^2}{2}}. \quad (6)$$

Then, by comparing it with the original power at each harmonic $S_w(k)$, the factor $\lambda(k)$ is formed,

$$\lambda(k) = \frac{L(k)}{\sqrt{\left|S_w(k)\right|^2 + L(k)^2}}, \quad (7)$$

which determines the division of power between the voiced and the unvoiced power estimates, $\hat{V}(k)$ and $\hat{U}(k)$ respectively:

$$\tilde{V}(k) = \begin{cases} \left(\sqrt{1-\lambda(k)^2}\right) \hat{V}(k) & \text{for } k \in \mathcal{H}, \\ \hat{V}(k) & \text{otherwise;} \end{cases} \quad (8)$$

$$\tilde{U}(k) = \begin{cases} \hat{U}(k) + \lambda(k)\hat{V}(k) & \text{for } k \in \mathcal{H}, \\ \hat{U}(k) & \text{otherwise.} \end{cases} \quad (9)$$

These modified spectra, just like those of the signal estimates (Eq. 5), are finally returned to the time domain by IDFT and windowed, resulting in the voiced and unvoiced power-interpolated signals, $\tilde{v}(n)$ and $\tilde{u}(n)$.

3. EVALUATION

The performance of the PSHF was evaluated using speech-like test signals $s(n)$, which were made up of a voiced part $v(n)$ and an unvoiced part $u(n)$:

$$s(n) = v(n) + u(n), \quad (10)$$

at sampling rate $f_s = 48\,\text{kHz}$. The voiced part was synthesized by convolving a pulse train $g(n)$, which was periodic at $f_0 = 130.8\,\text{Hz}$, with an appropriate impulse-response filter q:

$$v = g * q, \quad (11)$$

where $*$ denotes convolution. The filter q was built using the linear prediction coefficients (LPC, autocorrelation, 50-pole) obtained from an adult male mid-vowel [ɑ] recording (details below). The pulses were perturbed from their nominal amplitude and pitch period by a specified amount of random shimmer (0 dB or 1 dB) and jitter (0 %, 0.5 % or 3 %), respectively [6]. The unvoiced signal was similarly created by convolving Gaussian white noise $d(n)$ (zero mean, unit variance) with the LPC filter:

$$u = A\, d * q, \quad (12)$$

and the gain A was adjusted to give the desired HNR, initially at four levels (∞, 20 dB, 10 dB and 5 dB), but two more (0 dB and -5 dB) were included in an additional experiment.

So, using the PSHF signal estimates $\hat{v}$ and $\hat{u}$, the changes in signal-to-noise ratio (SNR), η_v and η_u, were calculated, as a measure of the decomposition algorithm's performance. The change in SNR for the harmonic component η_v is defined as the ratio of the unvoiced part's mean power to that of the residual error; conversely, the anharmonic performance η_u is the ratio

of voiced to error power. Both are expressed in decibels:

$$\eta_v = 10\log_{10}\left[\langle u^2\rangle/\langle e^2\rangle\right], \quad (13)$$

$$\eta_u = 10\log_{10}\left[\langle v^2\rangle/\langle e^2\rangle\right], \quad (14)$$

where the residual error is $e = (\hat{v} - v) = -(\hat{u} - u)$.

J	S	Harmonics-to-noise ratio (dB)					
%	dB	∞	20	10	5	0	-5
*0	0	$-,73$	$5,25$	$5,15$	$5,10$	$5,5$	$5,0$
0	0	$-,54$	$6,25$	$5,15$	$5,10$	–	–
	1	$-,22$	$1,20$	$5,14$	$5,10$	–	–
0.5	0	$-,28$	$4,24$	$5,15$	$5,10$	–	–
	1	$-,20$	$-1,18$	$4,14$	$5,10$	–	–
3	0	$-,13$	$-6,14$	$3,12$	$4,8$	–	–
	1	$-,14$	$-6,14$	$0,9$	$3,8$	–	–

Table 1: Harmonic and anharmonic performance of the PSHF (η_v, η_u in dB) versus target values of Jitter, Shimmer and HNR. (* Results obtained using $f_0 = 120.0$ Hz.)

The results, presented in Table 3, show that, for perturbation and noise levels normally found in speech, the PSHF enhanced the voiced component by 4–5 dB; the unvoiced component generally showed much greater improvements, of approximately 5 dB above the initial HNR. For typical HNRs (-5 dB to 20 dB), performance was unaffected by f_0; for other HNRs, minor changes in the random noise and quantization can have an exaggerated effect on the variance of measurements. Still, benefits to 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) tended to have a larger effect on performance than amplitude fluctuations (shimmer), for the range of values tested.

4. RESULTS

An adult male (PJ), a native speaker of British English RP, recorded a speech corpus that included sustained vowels (V = /ɑ, i, u/) and sustained fricatives in the form /ɑF:/ (F = /s, z/). 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–22 kHz band-pass, linear filter). It was recorded on DAT (Sony TCD-D7, $f_s = 48$ kHz), from which it was transfered digitally to computer.[1]

[1] Calibration tones were recorded to give an absolute reference to pressure, and background noise was recorded to assess the measurement-error floor.

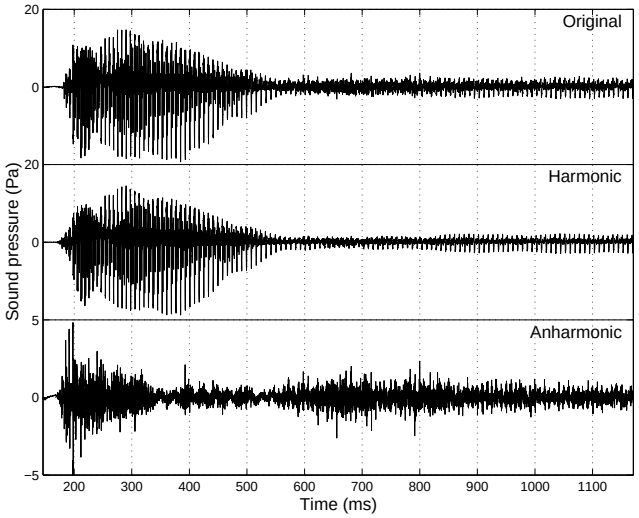

Figure 2: Time series from [ɑz:] by an adult male (PJ) of the original signal $s(n)$ (top), the harmonic component $\hat{v}(n)$ (middle), and the anharmonic component $\hat{v}(n)$ (bottom, quadruple amplitude scale).

Figure 2 illustrates the result of applying the PSHF to the utterance [ɑz:]. The majority of the signal energy is modeled by the harmonic or voiced component $\hat{v}$, which begins with a rapid growth of voicing that is then sustained at a high level during the vowel. After 200 ms, it starts to fade as the transition is made into the fricative, which appears to achieve a steady state from c. 560 ms onwards. The anharmonic or unvoiced component $\hat{u}$ is of a much lower amplitude in the vowel, although magnified four times in the graph, and follows a very different pattern: loudest initially, it quickly decays to its minimum in the latter part of the vowel, and reverts to an intermediate magnitude for the fricative, which reduces gradually. The signal $\hat{u}$ is noisy, in contrast to $\hat{v}$ which exhibits a regular pulsing throughout. Each of these general characteristics is as expected, including the initial surge of unvoiced noise, which could be generated by increased airflow at voice onset, although irregularities in phonation would also contribute some spurious elements (as indicated by the tests with synthetic signals).

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:

$$P_y(p) = \frac{\sum_{m=0}^{M-1} x(m)^2 y(p+m-M/2)^2}{\sum_{m=0}^{M-1} x(m)^2}, \quad (15)$$

using a smoothing window $x(m)$, which was set to a fixed length ($M = \langle N \rangle$, where $\langle\,\rangle$ denotes the time-average). The STP of the voiced component P_v, and

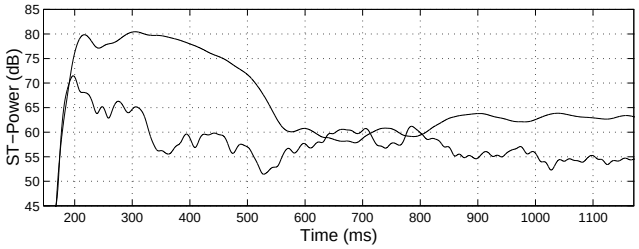

Figure 3: Short-time power (32 ms, Hanning window) of the voiced (thick) and unvoiced (thin) components, P_v and P_u, during [ɑz:].

that of the unvoiced component P_u, are plotted in Figure 3. Their trajectories agree with our earlier observations, but there is evidence of overshoot in the fricative (630–800 ms) before the final equilibrium was reached at c. 860 ms.

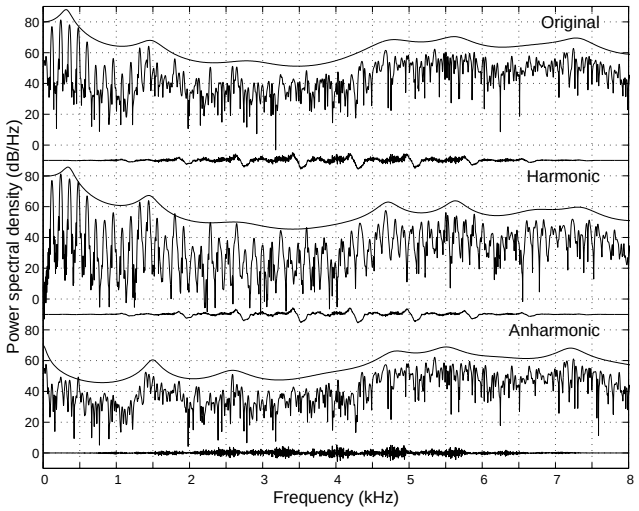

Figure 4: Power spectral density (85 ms, Hanning window centered at 900 ms, ×4 zero-padded, re. 2×10^{-5} Pa) computed from the original signal $s(n)$ (top) for the sustained fricative [z:] by an adult male subject (PJ), from the harmonic estimate $\tilde{v}(n)$ (middle) and from the anharmonic estimate $\tilde{u}(n)$ (bottom), whose time series are inset underneath each graph (anharmonic signal double scale).

Short sections of the signals around 900 ms were used to produce power spectra for the original and the two power estimates, $\tilde{v}(n)$ and $\tilde{u}(n)$. The spectra, each overlaid with a 50-pole LPC analysis, are plotted in Figure 4 (with their time waveforms). The waveforms show how the voiced signal $\tilde{v}$ has been purged of noise, which arises in the anharmonic part $\tilde{u}$ as pitch-synchronous packets. Most of the energy in the original spectrum comes in the first five harmonics but, even though the spectrum becomes more noisy at higher frequencies, there is a significant proportion in the range 4–8 kHz. However, the voiced spectrum maintains its periodic structure over all the frequencies plotted, while the unvoiced spectrum, being pervasively noisy, is devoid of harmonics. Although the smoothed LPC spectra display many similarities, there are notable differences in the resonance frequencies (e.g., peaks differ by 50 Hz at F_2, by 200 Hz at F_3). Moreover, the first formant F_1 is absent from the anharmonic curve, where their relative amplitudes are >30 dB apart, which is compatible with the net low-frequency anti-resonance excited by a frication source. At higher frequencies the anharmonic component dominates, also as expected.

5. CONCLUSION

We have presented a signal decomposition technique, its evaluation using synthetic signals, and results from its application to real speech. The potential for using the PSHF to enable separate analyses of voiced and unvoiced components in mixed-source speech was demonstrated. The PSHF gives the best decomposition during sustained phonation, since it is based on a harmonic model. Although jitter and shimmer in real speech can produce artefacts in the unvoiced component, tests indicated consistent improvements under typical conditions, suggesting that the algorithm can significantly aid the study of unvoiced sound production mechanisms, and the characterization of turbulence noise sources, such as frication and aspiration.

6. REFERENCES

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

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

[3] 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.

[4] P.J.B. Jackson and C.H. Shadle. Pitch-synchronous decomposition of mixed-source speech signals. *Proc. Joint Int. Cong. on Acoust. and Acoust. Soc. Am.*, Seattle, WA, 1:263–264, 1998.

[5] P.J.B. Jackson and C.H. Shadle. Decomposing speech signals into their simultaneous voiced and unvoiced components. *IEEE Trans. SAP*, submitted Apr. 1999.

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