Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech

Philip J.B. Jackson, *Member, IEEE*, and Christine H. Shadle, *Member, IEEE*

Manuscript received May 1999, revised March 2000 and Nov 2000. This work was supported by the Faculty of Engineering and Applied Science and the Department of Electronics and Computer Science, University of Southampton, UK. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yunxin Zhao.

P. Jackson is with the School of Electronic and Electrical Engineering,University of Birmingham, UK (e-mail: p.jackson@bham.ac.uk).

C. Shadle is with the Department of Electronics and Computer Science, University of Southampton, UK (e-mail: chs@ecs.soton.ac.uk).

Abstract

Almost all speech contains simultaneous contributions from more than one acoustic source within the speaker's vocal tract. In this paper we propose a method — the pitch-scaled harmonic filter (PSHF) — which aims to separate the voiced and turbulence-noise components of the speech signal during phonation, based on a maximum likelihood approach. The PSHF outputs periodic and aperiodic components that are estimates of the respective contributions of the different types of acoustic source. It produces four reconstructed time series signals by decomposing the original speech signal, first, according to amplitude, and then according to power of the Fourier coefficients. Thus, one pair of periodic and aperiodic signals is optimized for subsequent time-series analysis, and another pair for spectral analysis. The performance of the PSHF algorithm was tested on synthetic signals, using three forms of disturbance (jitter, shimmer and additive noise), and the results were used to predict the performance on real speech. Processing recorded speech examples elicited latent features from the signals, demonstrating the PSHF's potential for analysis of mixed-source speech. [EDICS: 1-ANLS]

Keywords

Periodic-aperiodic decomposition, speech modification, speech pre-processing.

Corresponding author: Dr. C. Shadle, Department of Electronics and Computer Science, University of Southampton, Southampton, Hants. SO17 1BJ, UK.
Tel: +44 (0)2380 592690, Fax: +44 (0)2380 594498, Email: chs@ecs.soton.ac.uk

I. Introduction

The acoustic cues that are central to our ability to perceive and recognize speech derive from a variety of acoustic mechanisms and are often classified according to the nature of the sound source: phonation, frication, plosion or aspiration [1], [2]. Identifying and characterizing the various sources is fundamental to speech production research [3–5], and to the classification of pathological speech. Recent studies of hoarse speech have concentrated on measures of roughness in phonation, e.g., [6], and yet turbulence-noise sources contribute largely to this effect (as breathiness). In normal or pathological speech, when more than one sound source is operating, it is difficult to segment the corresponding acoustic features, which typically overlap both in time and frequency, thus hindering the isolation of individual source mechanisms, and making it practically impossible to examine source interactions in any detail. Our particular area of interest is that of turbulence-noise sources in the vocal tract, and in order to explore these phenomena, we would like to be able to analyze the voiced and turbulence-noise components of mixed-source speech separately, possibly even to distinguish between all the different acoustic contributions. To that end we have developed a signal analysis technique for separating separating the periodic component, an estimate of thepart attributable to voicing, from the aperiodic component, an estimate of the part attributable to the simultaneous turbulence-noise source(s). Assessing the relative contribution of these two components as a harmonics-to-noise ratio (HNR) has long been a useful tool in the laboratory and the clinic [7–15], but there has been growing interest in more complete descriptions of the periodic and aperiodic signal components. Recent development of decomposition algorithms has been fueled by the demands of numerous speech applications: enhancement [16–21], modification [22–24], coding [25] and analysis [26], [27].

Decomposition is generally achieved by first modeling voicing deterministically, since voicing tends to be the larger signal component, and then attributing the residue to the estimate of the aperiodic component. Concentrating the periodic component into a certain region of a transformed space improves estimation of the model's parameters. The extraction of energy concentrations from the transformed signal is equivalent to the separation of deterministic and stochastic elements, which may be realized by a threshold operation, as in [28] using wavelets. Serra and Smith [25] combined peak-picking and tracking to code the voiced (deterministic) part and fitted line segments to the residual noise spectrum. However, the regularity of vocal fold vibration can be used to define the region of concentration, and to design a comb filter that

effectively averages successive pitch periods. The two main approaches are time domain (TD) and frequency domain (FD), although most contain elements of both.

TD models typically assume that noise is added to pulsed excitation of a time-varying, linear filter. One TD method is the comb filter with teeth periodically aligned on the pitch pulses. In order to adapt the spacing of the teeth of the comb filter in synchrony with variations in voicing, knowledge of the glottal pulse epochs is required. There have been many TD realizations of this pitch-synchronous method, which have accommodated timing variations by truncation and zero-padding [7], [29], [30], scaling [15], least-squares alignment [27], [31] or dynamic time warping [17].

FD methods estimate the Fourier series of pitch harmonics from the short-time Fourier transform (STFT), using the fundamental frequency f_0 to identify regions of the spectrum that correspond to voicing. Thus, they model voicing by a short-time harmonic series, whose parameters tend to be smoothed between analysis frames [16], [18], [20], [22], [23], [32], [33]. Laroche et al. [22] included linear f_0 variation within a frame, but in their example (pitch-synchronous, two-period window) the data were over-fitted, resulting in 3 kHz low- and high-pass filtered speech signals to represent the periodic and aperiodic components, respectively. Griffin and Lim [33] used the pitch harmonics to sub-divide the spectrum, and made a "voiced/unvoiced" decision on each harmonic band for coding the speech signal.

A compromise was proposed by de Krom [9], who created a harmonic comb filter in the FD using the rahmonics of the real cepstrum, which has been the basis for various implementations [9, 12–14, 35]. The log-spectrum obtained in this way from the rahmonic cepstrum (with the spectral envelope removed), which oscillates about zero, was then thresholded: frequencies for which it was greater than zero were defined as periodic, and those less than zero as aperiodic. Hence, the partitioning of regions in the cepstral domain provided a means of labelling those regions in the STFT spectrum.

For HNR estimation and synthesis applications (coding, copy-synthesis, modification), the accuracy with which the component signal is estimated is not important provided the salient signal properties are captured, which is also the case for certain types of analysis. More generally, though, we would like to analyze all the information that is known, without introducing inappropriate assumptions, and therefore provide an output with a minimum of distortion. After subtraction of the periodic model from the original spectrum, the residue's spectrum typically

lacks data at the harmonics, i.e., the region where voicing was concentrated, and values of zero may be the best estimate available for the aperiodic signal spectrum. Yet, for feature extraction from the power spectrum (e.g., for generating a stochastic model that reproduces the longer-term spectral characteristics of the aperiodic component), filling those gaps can be advantageous. Thus, spectral interpolation has been performed by linear prediction [22], and by approximating the spectral envelope with line segments [25] or cepstral coefficients [23]. One recently-published technique [14] uses a reconstruction algorithm, but we have discovered certain problems with it, which are described in the Appendix. Yet, we have followed a similar methodology in evaluating our algorithm.

Still, choosing a technique for one's own data and purpose is not straightforward. Lim et al. [30] showed that TD comb filtering decreased intelligibility, whereas a harmonic method increased it [18]. On the other hand, Qi and Hillman [12] found that an adaptation of de Krom's method performed poorly compared to another TD method [7]. Furthermore, it depends on one's objective and the particular kinds of speech one wishes to study. In our case, we are interested in sounds with a significant noisy element, such as voiced fricatives, where the voicing tends to be weak and pitch epochs are hard to identify precisely. This scenario would favor an FD approach, but even modal vowels are suitable candidates for FD decomposition if one wants to examine the spectral characteristics. TD methods, on the other hand, might be more appropriate at abrupt transitions in voicing, e.g., at onset.

Our technique, presented in the next section, is an FD method called the pitch-scaled harmonic filter (PSHF). It provides outputs that constitute our best estimate of the voiced and turbulence-noise signals (suitable for TD analysis), and spectrally-interpolated outputs that provide a better estimate of the components' power spectrum (suitable for power spectral analysis and modeling). Previous techniques have failed to distinguish these two objectives of the decomposition task. In Section III, the behavior and performance of the PSHF algorithm was tested using synthetic speech signals that contained three kinds of disturbance: shimmer (perturbed amplitude), jitter (perturbed fundamental frequency f_0), and additive Gaussian noise with variable burst duration. Section IV gives examples from speech recordings that were analyzed to illustrate some of the decomposition technique's capability, and Section V concludes.

II. Pitch-scaled harmonic filter

A. Basis for a pitch-scaled approach

We use the term pitch-scaled to refer to an analysis frame that contains a small integer multiple of pitch periods. It implies, for a constant sampling rate f_s, that the number of sample points in the frame N will be inversely proportional to the fundamental frequency f_0. This property complicates the windowing and re-splicing processes, but also brings substantial benefits: mainly that the harmonics of f_0 will be aligned with certain bins of the STFT (assuming we know the value of f_0). For example, if our analysis frame contains b pitch periods, then the frequency of the nbth Fourier coefficient will correspond to nf_0. When the frequency in question is not exactly aligned with one of the discrete frequency bins, leakage and spectral smearing take place, which produce errors in the form of bias.

For a single infinite sinusoid of frequency f_1 in Gaussian white noise (GWN), the highest peak in the DFT spectrum provides the least-squares estimate (minimum mean-squared error) of the magnitude, frequency and phase of the sinusoid, given enough samples are taken at a high enough rate [36], [37],[1] and coincides with the maximum likelihood estimate for the Gaussian distribution [38]. If f_1 is of the same order as the frequency resolution (i.e., $\leq 2f_s/N$), the negative-frequency image centered at $-f_1$ will not be sufficiently separated from it, and will bias the estimates [16], [36]. In contrast, if the analysis frame is chosen to have several whole cycles (with adequate f_s), f_1 will lie on a DFT bin, and the bias terms from interference and spectral leakage will disappear; the remaining error is unbiased Gaussian noise whose variance is proportional to that of the additive noise. When there is more than one sinusoid present in GWN, they must be sufficiently separated in frequency ($\delta f \geq 4f_s/N$) to maintain optimal (maximum likelihood) estimation of the deterministic components, as well as each meeting the earlier constraints [38], [39]. Again, these biases are avoided when N is scaled to the frequency of both sinusoids, which must therefore be harmonically related.

However, speech signals, although predominantly harmonic, are not composed of pure sinusoids of infinite duration. Vibration of the vocal folds tends to generate sound pressure signals that are approximately periodic, but whose amplitude and fundamental frequency fluctuate during voicing and change dramatically at voice onset/offset. Although some of the techniques we have

[1] Having too few samples would not give sufficient frequency resolution, and too low a sampling rate would provoke aliasing problems.

mentioned effectively applied a rectangular window, most used a smooth function, viz. Hanning or Hamming, to accommodate such non-stationarity. We have chosen to use a Hanning window, which still yields unbiased estimates when pitch-scaled, though it increases the variance of the error by 50 % [39], [40]. This step greatly enhances the technique's robustness to minor pertubations in periodicity. Cross-term bias errors between harmonics caused by deviations from perfect periodicity are reduced by a factor of 15 at the adjacent harmonic by the Hanning window, in comparison to a rectangular window (i.e., 24 dB, four bins away), as shown in Figure 1. Also, the half-power bandwidth of the main peak is increased from 0.44 bins to 0.72 bins at each harmonic, an increase of 60 %. Thus, despite being based on a maximum likelihood approach for estimating harmonically-related sinusoids, some of the idealized performance has been compromised to make the process more suitable for time-varying signals.

B. Overview

The pitch-scaled harmonic filter (PSHF), derived from a measure of HNR [8], was designed to separate the periodic and aperiodic components of speech signals. It is assumed that these components will be representative of the vocal-tract filtered voice source and noise source(s), respectively. The original speech signal $s(n)$ is decomposed primarily into the periodic (estimate of voiced) and aperiodic (estimate of turbulence-noise) components, $\hat{v}(n)$ and $\hat{u}(n)$ respectively. Further periodic and aperiodic estimates, $\tilde{v}(n)$ and $\tilde{u}(n)$, are computed based on interpolation of the aperiodic spectrum, which improves the spectral composition of the signals when considering features over a longer time-frame.

In the process of estimating the HNR from a short section of speech $s(n)$, Muta *et al.* [8] used the spectral properties of an analysis frame that was scaled to the pitch period in order to distinguish parts of the spectrum containing harmonic energy from those without. Hence, they applied a window function w of length $N(p)$ to $s(n)$, centered at time p, to form $s_w(n) = w(n)\, s(n + p - N/2)$. They computed the spectrum $S_w(k)$ by discrete Fourier transform (DFT) using a value of $N = b\tau$ that was a whole number b of pitch periods of length τ (in samples):

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

which concentrated the periodic part of s_w into the set of harmonic bins B, where B contains every bth coefficient: $\{b, 2b, 3b, \ldots, b(N-1)\}$. Choosing a four-pitch-period Hanning window

($b = 4$): $w(n) = 0.5\,(1 - \cos 2\pi n/N)$ for $n \in \{0, 1, \ldots, (N-1)\}$, the harmonics were translated to bins $\{4, 8, 12, \ldots\}$, while the bins halfway between $\{2, 6, 10, \ldots\}$ were kept free from spectral leakage of the periodic component. Thus, for an adult male speaker with pitch period of 8 ms ($f_0 = 125$ Hz), a 32 ms window would be used.

We have extended the process [41] to yield a full decomposition into periodic and aperiodic complex spectra, which can be converted back into time series, $\hat{v}$ and $\hat{u}$ respectively, as explained below. We also propose an interpolation step for improving power-spectral estimation, which produces a further pair of signals $\tilde{v}$ and $\tilde{u}$. The outputs can later be analyzed using any standard technique: $\hat{v}$ and $\hat{u}$ for TD analysis, $\tilde{v}$ and $\tilde{u}$ for FD analysis. For time-frequency analysis, we define a threshold 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. The remainder of this section describes the Muta et al. [8] pitch estimator, the segmentation of speech signals into frames and the PSHF algorithm.

C. Pitch estimation

The PSHF relies on the window length being scaled to match the time-varying pitch period: $N(p) = b\tau(p)$. The pitch-tracking algorithm estimates the period τ by sharpening the spectrum at the first H harmonics. The sharpness is described in terms of the higher and lower spectral spread, S_h^+ and S_h^- respectively, which are defined for a given window at each harmonic, $h \in \{1, 2, \cdots, H\}$ as:

$$S_h^+(N,p) = |S_w(bh+1)|^2 - \frac{|S_w(bh)|^2}{|W(h\,\Delta f_0)|^2}\left|W\left(h\,\Delta f_0 - \frac{1}{N}\right)\right|^2 \quad (2)$$

$$S_h^-(N,p) = |S_w(bh-1)|^2 - \frac{|S_w(bh)|^2}{|W(h\,\Delta f_0)|^2}\left|W\left(h\,\Delta f_0 + \frac{1}{N}\right)\right|^2, \quad (3)$$

where $\Delta f_0 = 1/\Delta\tau = b f_s/\Delta N$,

$$W(k) = \frac{N}{2}\left(\operatorname{sinc}\pi kN + \frac{1}{2}\left[\operatorname{sinc}\pi(kN-1) + \operatorname{sinc}\pi(kN+1)\right]\right)\exp -j\pi\Delta f_0 N\,,$$

for the Hanning window, and $\operatorname{sinc} x = \sin(x)/x$. Thus, the spectral smearing due to the window is calculated for the higher and lower bins adjacent to each harmonic, $k = bh \pm 1$, and the values are compared to the measured values in those bins. The optimum pitch estimate $N_{\text{opt}}(p)$ is obtained by minimizing the difference between the calculated and measured smearing in a

minimum mean-squared error sense, according to the cost function at time p:

$$J(N,p) = \sum_{h=1}^{H} \left(S_h^+(N,p)^2 + S_h^-(N,p)^2 \right) ; \quad (4)$$

see [8] for further details. The optimization is perfectly matched to the PSHF because, using the same window, it maximizes the concentration of signal energy into the harmonic bins.

For each section of voiced speech, the initial estimate of $N(p)$ was set manually. For larger data sets, standard methods could easily be implemented for automatic initialization, e.g., [42–44]. The pitch tracker operated as follows: (i) window speech signal (N-point, Hanning); (ii) evaluate cost function $J(N,p)$ near current estimate $N(p)$; (iii) update the current estimate to N_{opt} (value at minimum cost); (iv) increment time p and repeat.

D. Windowing and re-splicing

Windowing was used in the PSHF not only to process the data in finite frames, but also to allow the piecewise stationary model to adapt in line with the many kinds of variation in the speech production system: amplitude, fundamental frequency, formant frequencies, voice onset/offset and other transients. After decomposing a frame, the output signals were recombined with the results of preceding frames by overlapping and adding.

For simplicity, the center positions p_i of the frames i were spaced at a constant interval: $\alpha = p_i - p_{i-1}$. However, since the window size was not generally constant, neither was the signal weighting; lower fundamental frequency regions, having longer windows $w_i(n)$, accrued more weighting than higher f_0 regions. Therefore, to normalize the final output signals, i.e., the re-spliced periodic and aperiodic components, they were multiplied by $W(m)$, the reciprocal of the sum of the contributions from the windows w_i:

$$W(m) = \frac{1}{\sum_i \{ w_i (m - p_i + N(p_i)/2) \}} , \quad (5)$$

for all frames i that included the point m (not necessarily contiguous). Alternatively, each frame's window could be normalized to give an even point-wise weighting, as done in [30]. A cosine ramp was applied to each end of the normalization factor $W(n)$ to fade out sections of voicing at onset and offset.

E. Algorithm

Harmonic filter. Let us consider how the PSHF algorithm performs the decomposition in the FD for a single frame, centered at time p. (Note: all functions within the algorithm are adaptive

and depend on p, but for clarity, we omit the argument p hereafter.) After applying the pitch-scaled Hanning window to the speech signal to get $s_w(n)$, the PSHF algorithm computes $S_w(k)$ by DFT, as depicted in Figure 2. The harmonic filter (HF) takes the pitch harmonics from S_w and doubles the coefficients to form the harmonic spectrum $\hat{V}(k)$, compensating for the mean window amplitude of 0.5:

$$\hat{V}(k) = \begin{cases} 2S_w(k) & \text{for } k \in B \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

where $B = \{b, 2b, \ldots, (N-1)b\}$. This spectrum, when returned to the time domain by inverse DFT (IDFT), produces a signal that is periodic with no envelope shaping, so these four pitch periods are windowed to yield the periodic signal estimate $\hat{v}_w(n)$:

$$\hat{v}_w(n) = \frac{w(n)}{N} \sum_{k=0}^{N-1} \hat{V}(k) \exp\left(j \frac{2\pi n k}{N}\right). \tag{7}$$

The aperiodic signal estimate is the difference between this and the input signal: $\hat{u}_w(n) = s_w(n) - \hat{v}_w(n)$. Alternatively, in the frequency domain, we can subtract $\hat{V}$ from the unwindowed spectrum:

$$\hat{U}(k) = \begin{cases} S(k) - 2S_w(k) & \text{for } k \in B \\ S(k) & \text{otherwise}, \end{cases} \tag{8}$$

and then the aperiodic component $\hat{u}_w$ comes from applying the IDFT and window, as before. As a result, any errors in the periodic estimate caused by the decomposition algorithm are (wrongly) attributed to the aperiodic signal. Note that the number of pitch periods b can potentially be any integer that achieves a harmonic concentration, viz. $b \in \{2, 3, 4, \ldots\}$. There is inevitably a trade-off between time and frequency resolution which, among other things, balances the noise rejection performance against the tolerance to jitter and shimmer. We have found that $b = 4$ offers a favorable compromise, but we have not tested alternatives.

Power interpolation. The spectrum of the estimated aperiodic signal $\hat{U}_w(k)$ contains gaps at the harmonics, where the coefficients are of zero amplitude, since $\hat{U}_w(k) = S_w(k) - (2S_w(k))/2 = 0$ for $k \in B$. However, subsequent analysis often involves computing power spectra or spectrograms, which depend on the squared magnitude of the Fourier coefficients, and the gaps therefore give strongly biased under-estimates. We can improve the power estimates by filling $\hat{U}_w$ in at the harmonics. If we assume that the aperiodic component is the result of a stochastic process with a smoothly varying frequency response, we would expect the power in any frequency bin

to be similar to its adjacent bins. Therefore, we calculate $L(k)$, a frequency-local estimate of $|U_w|$ at the harmonics, by power interpolation (PI) of the values of the aperiodic spectrum in the adjacent bins, $\hat{U}_w(k \pm 1)$:

$$L(k) = \sqrt{\frac{\left|\hat{U}_w(k-1)\right|^2 + \left|\hat{U}_w(k+1)\right|^2}{2}} \quad \text{for } k \in B\,. \tag{9}$$

The RMS amplitude $L(k)$ is compared with the periodic spectrum $\hat{V}_w(k) = S_w(k)$ for $k \in B$, to determine the real factor $\lambda(k)$, which is the proportion of the coefficient to be allocated to the revised aperiodic estimate $\tilde{U}(k)$, for each harmonic:

$$\lambda(k) = \frac{L(k)}{\sqrt{|S_w(k)|^2 + L(k)^2}}\,. \tag{10}$$

The remainder of the power is left with the revised harmonic estimate $\tilde{V}(k)$, so we have:

$$\tilde{V}(k) = \begin{cases} \sqrt{1-\lambda(k)^2}\,\hat{V}(k) & \text{for } k \in B, \\ \hat{V}(k) & \text{otherwise;} \end{cases} \tag{11}$$

$$\tilde{U}(k) = \begin{cases} \hat{U}(k) + \lambda(k)\hat{V}(k) & \text{for } k \in B, \\ \hat{U}(k) & \text{otherwise.} \end{cases} \tag{12}$$

Hence, by using the original phase information for both components, $\arg(S_w(k))$, we can reconstruct the power-based time series $\tilde{v}_w(n)$ and $\tilde{u}_w(n)$ in a way that is consistent between overlapping frames. These signals retain the detail of the original time series, while avoiding misleading artefacts in the power spectrum in the form of troughs or valleys at the harmonics, and thus are suitable for long-term spectral analysis. As shown in Figure 2, the algorithm generates four complex spectra, $\hat{V}(k)$, $\hat{U}(k)$, $\tilde{V}(k)$ and $\tilde{U}(k)$, from a single input. After inverse-transforming and windowing, these are output as four time-series signals: $\hat{v}_w(n)$, $\hat{u}_w(n)$, $\tilde{v}_w(n)$ and $\tilde{u}_w(n)$, respectively. Each of these can be combined with the outputs from previous frames by sequential overlapping and adding to reconstruct two pairs of complete signals corresponding to the original signal $s(n)$: the periodic and aperiodic signal estimates $\hat{v}(n)$ and $\hat{u}(n)$, and the periodic and aperiodic power-based estimates $\tilde{v}(n)$ and $\tilde{u}(n)$.

III. Testing

A. Signal generation

The PSHF was tested with synthetic speech-like signals and the accuracy of its decomposition evaluated. The signals $s(n)$ were generated in the TD (avoiding any potential artefacts from

later FD filtering) by convolving excitation signals $c(n)$ with an appropriate filter $q(n)$:

$$s(n) = c(n) * q(n)\,. \tag{13}$$

Each excitation signal $c(n)$ was the sum of a pulse train $g(n)$ (with samples $\ldots 0,\ 1,\ 0,\ 0,\ \ldots$) and GWN $d(n)$:

$$c(n) = g(n) + d(n)\,. \tag{14}$$

The pitch period and amplitude of $g(n)$ were perturbed from their nominal values ($\bar{f}_0 = 120$, 130.8 or 200 Hz, $\bar{a} = 1$) by specified degrees of jitter (0, 0.25, 0.5, 1 or 3 %) and shimmer (0, 0.5, 1 or 1.5 dB), respectively.[2] Normal values for jitter and shimmer during modal phonation are typically less than 0.7 % and 0.5 dB, respectively [45] (less than 1 % and 0.25 dB according to [46]), although they can be as much as 3 % and 1 dB [11]. The noise, $d(n)$, was added at six levels with HNRs of ∞, 20, 10, 5, 0 or -5 dB. In some cases, the amplitude of the noise was modulated by a rectangular wave in time with the pulses to give a burst duration 60 % of the pitch period.

A set of linear predictive coding (LPC) coefficients (50-pole, autocorrelation) was computed for a male [ɑ], using a section from the middle of the first vowel in a recorded nonsense word (see Section IV-B for details). Each excitation signal, $c(n)$, was passed through the corresponding LPC synthesis filter, $q(n)$, at sampling rate of 48 kHz.

B. Parameters

Jitter is a measure of fluctuation in the pitch period (or fundamental frequency) of the voice. Usually expressed as a percentage, it is defined [47–49] as:

$$\hat{\sigma}_T = \frac{\mathrm{E}\left[|\tau_i - \tau_{i-1}|\right]}{\mathrm{E}\left[\tau_i\right]} \times 100\ (\%)\,, \tag{15}$$

where the period of the ith pulse, $\tau_i = t_i - t_{i-1}$, is the difference between the current pitch epoch t_i and the previous one, and $\mathrm{E}[\ ]$ denotes the expected value. It can be evaluated for all pulses in a given section of signal, or restricted to a region of that signal, to give a more time-specific measurement.

[2]The jitter and shimmer perturbations created respectively by Eqs. 16 and 19 do not necessarily represent realistic patterns of f_0 variation, but are used to illustrate the effect of perturbations on the PSHF. The fine time resolution of the PSHF leaves it unaffected by low-frequency perturbations, such as vibrato, but the above test methodology provides quantitative and self-consistent results.

For generating signals, each specified jitter value was used to modify the period [11]:

$$\tau_i = \frac{1}{\bar{f}_0}\left(1 + \frac{r_i\sqrt{\pi}}{2}\frac{\sigma_T}{100}\right), \tag{16}$$

where $\bar{f}_0$ is the nominal fundamental frequency and r_i is a random variable with a Gaussian distribution of zero mean and unit standard deviation. The factor of $\sqrt{\pi}/2$ is needed to match the standard deviation of τ_i to the mean difference between two such variables, $|\tau_i - \tau_{i-1}|$.

In real speech, the jitter and equilibrium fundamental frequency vary with time. So, using a window function $x(n)$ (e.g., triangular, Hanning, Hamming, Kaiser, etc.) offers a means to evaluate the short-time jitter:

$$\tilde{\sigma}_T(p) = \frac{\langle |\tau_i - \tau_{i-1}|\ x(t_i - p)\rangle}{\langle \tau_i\, x(t_i - p)\rangle} \times 100\ (\%), \tag{17}$$

in the vicinity of point p, where $\langle\ \rangle$ denotes the time average. Note that, in practice, computation of Eq. 15 over a finite number of pitch periods is equivalent to Eq. 17, when $x(n)$ is rectangular. To identify the pitch instants, τ_i, we used zero-crossing [10] and peak-picking [50] methods to refine initial manual estimates.

Shimmer is a measure of the fluctuation of the amplitude of the voice. Usually expressed in decibels, it is defined [46], [48] as:

$$\hat{\sigma}_A = 20\log_{10}\left(\frac{\mathrm{E}\left[|a_i - a_{i-1}|\right]}{\mathrm{E}\left[a_i\right]}\right)\ (\mathrm{dB}), \tag{18}$$

where a_i is the amplitude of the ith pulse. For generating signals, the pulse amplitude was calculated as [11]:

$$a_i = \bar{a}\left(1 + \frac{r_i\sqrt{\pi}}{2}10^{0.05\sigma_A}\right), \tag{19}$$

and the corresponding short-time shimmer was:

$$\tilde{\sigma}_A(p) = 20\log_{10}\left(\frac{\langle |a_i - a_{i-1}|\ x(t_i - p)\rangle}{\langle a_i\, x(t_i - p)\rangle}\right)\ (\mathrm{dB}). \tag{20}$$

For real speech, each pulse amplitude, a_i, was estimated using the RMS amplitude of the signal, windowed by an asymmetric Hanning window, extending one pitch period either side of the pitch instant in question.

The HNR is often used as a measure of the relative amplitudes of the voiced and noise components and is defined [30], [48] as:

$$\hat{\sigma}_N = 10\log_{10}\left(\frac{\mathrm{E}\left[v^2\right]}{\mathrm{E}\left[u^2\right]}\right)\ (\mathrm{dB}). \tag{21}$$

For the synthetic signals, the gain of the noise signal, $d(n)$, was adjusted relative to that of the pulse train, $g(n)$, to give the desired ratio σ_N. The short-time HNR, based on the periodic and aperiodic estimates, is:

$$\tilde{\sigma}_N(p) = 10\log_{10}\left(\frac{\langle \tilde{v}^2(n)\, x^2(n-p)\rangle}{\langle \tilde{u}^2(n)\, x^2(n-p)\rangle}\right) \text{(dB)}\,. \tag{22}$$

C. Performance calculation

As a result of decomposition of the speech $s(n)$, we want a periodic signal $\hat{v}(n)$ that represents the best estimate of the voiced component, defined as having the minimum mean squared error between the actual voiced component time series $v(n)$ and the estimate $\hat{v}(n)$. Similarly, we want the aperiodic signal $\hat{u}(n)$ to be the best estimate of the additive noise $u(n)$. The error $e(n)$, defined as $e = \hat{v} - v = -(\hat{u} - u)$, is equally (and oppositely) present in the periodic and aperiodic components.

The performance of the PSHF was assessed by considering the change in signal-to-error ratio (SER) for each component. The jitter and shimmer perturbations of the pulse train were considered intrinsic to the synthetic voicing signal, whereas the additive noise was treated as the product of another (turbulence-noise) source, and thus attributed to the aperiodic component. Therefore, for the periodic component, the additive noise was the initial 'error' on the voiced component, the 'signal'. Conversely, for the aperiodic component, the actual voiced component was taken to be the initial 'error' on the additive-noise 'signal'. Hence, the periodic performance η_v and the aperiodic performance η_u are:

$$\eta_v = 10\log_{10}\left(\frac{\langle v^2\rangle/\langle e^2\rangle}{\langle v^2\rangle/\langle u^2\rangle}\right) = 10\log_{10}\left(\frac{\langle u^2\rangle}{\langle e^2\rangle}\right) \text{(dB)}\,, \text{ and} \tag{23}$$

$$\eta_u = 10\log_{10}\left(\frac{\langle v^2\rangle}{\langle e^2\rangle}\right) \text{(dB)}\,. \tag{24}$$

It follows that evaluating the change in SER for the periodic and aperiodic estimates from the synthetic speech constitutes a more rigorous performance metric for reconstructing signals than a comparison of prescribed HNR (before synthesis) versus measured HNR (after decomposition). So, although we include some HNR measurements to aid comparison with other algorithms, we prefer to use the SER to describe the performance of the PSHF.

D. Results

First, the cost function $J(N,p)$ was used by the pitch tracker ($H = 8$ harmonics) to optimize the window length $N(p)$ for each synthetic signal. The signals were then decomposed by the PSHF algorithm into periodic and aperiodic components, $\hat{v}$ and $\hat{u}$ respectively, the estimates of the voiced and turbulence-noise parts. For this study, we were deliberately conservative, centering frames on every sample point (offset $\alpha = 1$), which was computationally expensive.

Figure 3 shows the results for three periodic signals corrupted by various amounts of either constant or modulated noise. The performance was positive in all but a few extreme cases, and was typically $\eta_v \approx 5$ dB for the periodic component and $\eta_u \approx \sigma_N + 5$ dB for the aperiodic one. For $\sigma_N \leq 0$ dB, the performance deteriorated and in some cases became negative; this deterioration was more pronounced for modulated noise. At infinite HNR ($\sigma_N = \infty$ dB), improvements in the aperiodic SER were 73, 54 and 50 dB respectively, for the three values of $\bar{f}_0$: 120, 130.8 and 200 Hz. Thus, pitch quantization and spectral smearing defined a performance limit by producing errors up to 1/300th of the original signal with no jitter, shimmer or noise disturbance.

The results were almost identical for all $\bar{f}_0$ values, a characteristic of pitch scaling, except at low HNRs where pitch tracking errors produced spurious readings. Similarly, altering the envelope of the noise, although perhaps making the tracker more error-prone, did not significantly affect the quality of the decomposition. In another study [51], we synthesized signals with constant-amplitude noise and noise modulated by f_0, and showed that the respective constant and modulated envelopes of the reconstructed noise signals were retained. These results suggest that any modulation observed in components of speech is real rather than a processing artefact.

Figure 4 illustrates the effects of jitter (left) and shimmer (right) on the PSHF performance, in combination with constant noise added at various levels. The trends are qualitatively similar for both perturbations. For example, when there is no noise, there is a notable performance degradation with the introduction of any jitter or shimmer. However, for the range of values chosen, fluctuations in the pitch period (jitter) have a larger effect on performance than amplitude fluctuations (shimmer). Where there is already one disturbance, i.e., HNRs of 20, 10 or 5 dB, the introduction of a second one, either jitter or shimmer, is less marked. The performances are generally positive, except for η_v at the higher levels of jitter ($\sigma_T \geq 1.5\,\%$) and shimmer ($\sigma_A \geq 1.5$ dB) with high HNR ($\sigma_N \geq 20$ dB), for which the initial error was relatively small. The grid of results in Table I extends this principle to the combination of all three disturbances, whose

worst element puts a bound on the performance. Indeed, the performance can even improve, as occurred for jitter of 3 % when shimmer was added. For normal speech, the presence of all three disturbances degrades performance by 1 to 2 dB with respect to the noise-only case (in Fig. 3).

Although not principally designed for such a purpose, the power-based outputs of the PSHF, $\tilde{v}$ and $\tilde{u}$, may be used as a measure of the total power of each component. Hence, by comparing $\langle \tilde{v}^2 \rangle$ with $\langle \tilde{u}^2 \rangle$, an estimate of the HNR may be formed, where $\langle\ \rangle$ denotes time averaging. The measured HNRs, calculated for the signals from Figure 3, are just above the true (prescribed) HNRs in all cases, except for $\sigma_N = \infty$ (the no-noise case discussed above), as shown in Figure 5. The measured HNRs varied little with $\bar{f}_0$, and the noise envelope (constant or modulated) had a negligible effect. The discrepancy between the measured and prescribed HNRs is largest for the cases with most tracking errors, i.e., at -5 dB, but otherwise it is ca. 1–2 dB. Note that the decomposition anomaly evident in Figure 3 ($\sigma_N = 0$ dB, modulated, $\bar{f}_0 = 130.8$ Hz) is not apparent in these results, because the measured HNR, which is the ratio of the component powers, is not based on the actual decomposed signals and merely compares their mean square values.

In summary, the introduction of any form of disturbance, from noise or perturbation, drastically reduced the performance from that under ideal conditions, but the PSHF continued to give robust performance in the presence of secondary or tertiary disturbances. For positive HNR values, the algorithm enhanced the aperiodic component (i.e., improved its SER) much more than the periodic one, which particularly aids us in the study of turbulence-noise components of mixed-source sounds. For recordings of normal speech, the results suggest improvements to the SER of a factor of about five for the aperiodic component ($\eta_u \approx 10$–20 dB for $5\,\text{dB} \leq \sigma_N \leq 15$ dB) and about two for the periodic component ($\eta_v \approx 4$ dB).

IV. Application to real speech

A. Recording method

Two adult, native speakers of British English RP, one male (PJ) and one female (SB), recorded a speech corpus containing nonsense words and sustained vowels (V = /ɑ, i, u/) in a sound-treated room. The sound pressure at 1 m was measured 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), and recorded onto DAT (Sony TCD-D7, $f_s = 48$ kHz). The 16-bit data were then transfered digitally to computer

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

B. Example 1: nonsense word

Our first example is the nonsense word [pʰɑzɑ] spoken by subject PJ. A decomposition of the entire word is illustrated in Figure 6 as two sets of spectrograms: wide-band using s, $\hat{v}$ and $\hat{u}$, and narrow-band using s, $\tilde{v}$ and $\tilde{u}$, respectively.[3] In the voiceless regions (0–10 ms and 720–800 ms), there was no need to extract the voiced component, so the PSHF was not applied. For our purposes the voiced/voiceless decision was made manually, although there are many ways to do so automatically (e.g., [42]). Therefore, the periodic outputs were set to zero, $\hat{v} = \tilde{v} = 0$, and the aperiodic outputs were set equal to the original signal, $\hat{u} = \tilde{u} = s$, during the voiceless periods at either end of the utterance.

In the wide-band spectrogram of the original signal (Fig. 6, top left), the main cues are visible: the burst stripe (at 10 ms) with subsequent aspiration noise and formant transitions; the onset of voicing (at 70 ms) exciting the formants, which continues until the start of the fricative (ca. 300 ms) when it begins to die down, F1 and F2 diverge, and the high frequency noise grows (until 380 ms); the second vowel (from 420 ms), and finally voice offset (at 720 ms). The periodic component $\hat{v}$ retains a small yet significant part of the frication noise, but generally the voicing stripes are cleaner and more pronounced. The aperiodic spectrogram is generally mottled in appearance, as is characteristic for noisy sounds. However, different frequency regions are excited in each of the four source types: burst (all frequencies simultaneously with lowered formants), aspiration (all frequencies), mid-vowel (principal formants), and frication (higher formants).

Vertical striations can be seen in the high-frequency turbulence noise during the onset of frication, which become less noticeable towards mid-fricative. There is some contamination from the voiced part, particularly in unsteady regions (i.e., 200 ms, 270 ms, 450 ms) and at voice onset (70–100 ms), which correspond to rapid changes of f_0 and local peaks in the cost function.

In the narrow-band spectrograms (Fig. 6, right), one can see fine horizontal striations from the harmonics of the fundamental frequency, both for the original signal and more obviously for the periodic component, persisting throughout phonation. Some prosodic effects are visible, such as when harmonics cross a formant (e.g., F3 at 2.7 kHz, 100–200ms). Again, the periodic spectrogram is cleaner than the original one, while the aperiodic one remains mottled. The horizontal

[3] This is consistent with the discussion in Section II-B.

stripes are evident in short sections of the aperiodic spectrogram, where voicing perturbations have caused some leakage. However, the overall structure of $\tilde{u}$ is not generally periodic: note the stripes are absent from the pulsed frication noise and from much of the vowel sections (e.g., 590–680 ms), while the wide-band spectrogram shows clear signs of modulation. This implies that the PSHF has extracted pulsed noise into the aperiodic estimate, which would most likely be from aspiration in the vowels.

Figure 7 gives an expanded view of the reconstructed signals at the vowel-fricative transition [-ɑz-]. In agreement with earlier observations [22], [52], the aperiodic component exhibits modulation by the voice source during development of the fricative (300–370 ms). The effect becomes negligible (around 380 ms) as voicing dies away and the noise level increases. The periodic pulses in $\hat{v}$ become less spiky, consistent with a weaker glottal closure, and approach the form of a simple harmonic oscillation (that is increasingly contaminated by the frication noise).[4]

C. Example 2: sustained vowel

Our second example, a sustained vowel [ɑː] produced by SB, was decomposed to give the periodic and aperiodic estimates, $\hat{v}$ and $\hat{u}$, and the power-based estimates, $\tilde{v}$ and $\tilde{u}$ respectively. Figure 8 depicts the spectra derived from s, $\tilde{v}$ and $\tilde{u}$, using a steady section from the center of the vowel.

The periodicity of $\tilde{v}$ is strongly marked by the harmonic peaks of its spectrum, still noticeable above 8 kHz. Reassuringly, the levels of the harmonic peaks remain practically untouched by the PSHF, while the inter-harmonic troughs were deepened. Both components show the effect of the principal formants, although their spectral tilts are very different. Apart from the very low-frequency noise ($f < 50$ Hz, mostly wind noise generated at the microphone), $\tilde{u}$ contains a much greater portion of the original signal at high frequencies ($f > 3$ kHz), as expected for flow-induced, turbulence noise. Moreover, in the detail, there are features distinct to the aperiodic spectrum, such as a peak which had been hidden between the first two harmonics (~ 250 Hz) and a trough just above F2 at 1.4 kHz.

The jitter, shimmer and HNR were measured locally for the same section of speech: $\tilde{\sigma}_T = 0.9$ %, $\tilde{\sigma}_A = 0.07$ dB and $\tilde{\sigma}_N = 14$ dB. These values were used to predict the PSHF's performance by

[4]It is possible to incorporate heuristic knowledge of speech signals to reduce the cross-contamination of the periodic component, e.g., by low-pass filtering [22], but subjective assessment indicates that additional processing often incurs a loss of intelligibility [30].

interpolating the results of Table I: $\eta_v \approx 3\,\text{dB}$ and $\eta_u \approx 17\,\text{dB}$. Thus, we can claim with some confidence that the periodic component is an improved estimate of the voiced part over the original signal, and that the majority of the aperiodic component was produced by a turbulence-noise source.

D. Summary

For the nonsense word (Ex. 1), we discussed spectrograms of the decomposed signals and used them to extract features in the individual components. Examination of the time series at the vowel-fricative transition revealed the weakening of modulation of the aperiodic part as the fricative developed. When one listens to the separated components, the periodic component of [p^hɑzɑ] sounds like [ɑzɑ] with less emphasis on the fricative, and the aperiodic component like a whispered version of the original, albeit with some remnants of voicing.

The PSHF provides separate output signals that can be analyzed individually for feature extraction [24], [53], or in tandem to investigate interactions of voicing and noise sources. Indeed, the PSHF has been used to enable us to examine the timing relationship between voicing and the modulation of frication in a number of voiced fricatives [51]. We have also used it to compare the aperiodic component of voiced phonemes with their voiceless correlates to evaluate differences in their production [41]. Both the performance predictions and the interpretations of the periodic and aperiodic spectra (e.g., Fig. 8) present a compelling argument for their validity.

V. Conclusion

An analysis technique has been developed for decomposing mixed-source speech signals that is based on a pitch-scaled, least-squares separation in the frequency domain. The PSHF technique provides estimates of the voiced and turbulence-noise components, as periodic and aperiodic parts, using only the speech signal. The components can subsequently be subjected to any standard analysis, as time series or as power spectra, for instance.

Tests on synthetic speech demonstrated the PSHF's ability to reconstruct the components, despite corruption by jitter, shimmer and additive noise. It achieved improvements to the SER of the periodic and aperiodic parts of $\eta_v = 5\,\text{dB}$ and $\eta_u = 15\,\text{dB}$, respectively, for typical speech conditions ($\sigma_T = 0.5\,\%$, $\sigma_A = 0\,\text{dB}$ and $\sigma_N = 10\,\text{dB}$). The performance decreased gradually with increased corruption over a normal range of test conditions. Processing real speech examples resulted in convincing decompositions that revealed features particular to the individual

components.[5] Local measurements of the perturbation of the original speech signal were then used to predict the accuracy of the decomposed signals as estimates of the voiced and turbulence-noise components.

The main limitations of the technique concern its computational efficiency and robustness of the pitch-tracker to deviations of the input speech signal from periodicity. The current implementation of the algorithm is far from real-time, although there is plenty of scope for reducing the amount of computation. Jitter, shimmer, transients and voice onset/offset transitions all tend to produce errors degrading performance, although a high degree of robustness has been demonstrated across normal speech conditions. Further work is needed to explore potential refinements to the PSHF, and to benchmark it against other TD and FD methods. However, there is potential for applying the PSHF to a variety of speech problems, particularly the analysis of mixed-source speech production and speech modification.

Acknowledgments

We would like to thank Drs. Bob Damper and Paul White, and two anonymous reviewers for their helpful comments on earlier versions of this manuscript, and Dr David Lopes and Prof. Yegnanarayana for helpful discussions. We are also grateful for the cooperation of Dr. d'Alessandro who furnished us with signals for comparison.

[5]Sound files can be found at the project web site [54].

Appendix

Periodic-aperiodic decomposition

The periodic-aperiodic decomposition (PAPD) algorithm is an alternative technique, which was developed by Yegnanarayana, d'Alessandro and Darsinos [14], [34], [35], [55] for separating the voiced and noise components of a mixed-source speech signal. The algorithm would appear to have the characteristics needed for our purposes, and we have indeed adopted aspects of their general approach. However, as mentioned in the Introduction, we have discovered certain problems with it, which we have used to inform the development of our PSHF. This critique summarizes their algorithm, argues that the interpolation procedure converges to the original signal, presents supporting simulation results and discusses their approach in general. For consistency of notation within this article, many of their symbols have been altered. The substitutions are given in Table II.

A. Précis

Figure 9 is a schematic summary of the PAPD, which illustrates the way the algorithm is encased by an LPC analysis/synthesis shell. This shell pre-whitens the input signals before decomposition and returns the spectral coloring (e.g., from the formants) afterwards. The algorithm operates on the excitation signal, $c(n)$, to separate the periodic and aperiodic components in a two-stage process.

The first stage makes an initial separation in the frequency domain using a cepstral filter. The signal $c(n)$ is windowed and zero-padded to form $c_w(n)$, where N is the DFT length and $(N/2-1)$ the window length. Its spectrum $C_w(k)$ and real cepstrum are computed. The periodic region of the cepstrum is partitioned by extracting the first rahmonic, in a manner similar to de Krom's [9], and the DFT is computed. By comparing the log-spectrum to zero, the bins of the spectrum $C_w(k)$ are assigned to either the periodic component (positive values) or the aperiodic component (negative values). The initial aperiodic estimate is thus set equal to the original spectrum for the aperiodic bins B_d and zero elsewhere:

$$D_0(k) = \begin{cases} C_w(k) & \text{for } k \in B_d, \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

The second stage is an iterative interpolation process (IIP), involving repeated transformations between FD and TD. The IDFT of $D_0(k)$ is not generally time-compact like $c_w(n)$: that is,

$d_0 \neq 0$ for $N/2 \leq n \leq N$. The interpolation sets these points to zero, computes the DFT, resets $D_m(k) = C_w(k)$ for $k \in B_d$, computes the IDFT and so on. Setting the points to zero is equivalent to multiplying by a rectangular window: $\xi(n) = 1$ for $n \in \{1, 2, \ldots, N/2 - 1\}$, 0 for $n \in \{N/2, \ldots, N\}$. The process is repeated for 20 iterations, which Yegnanarayana et al. considered enough to allow $D_m(k)$ to converge.

Their results of decomposing synthetic signals show a strong correlation between the HNR that was prescribed when generating the synthetic speech (prescribed HNR), and the value calculated from the decomposed signals (measured HNR), which they called the periodic-aperiodic energy ratio. However, there appears to be a tendency to under-estimate the aperiodic component, since all reported values of measured HNR were too high, except in the total absence of noise. The effects of jitter, shimmer and f_0 glides are also highly significant, producing a large reduction in the measured HNR; a normal degree of jitter ($\sim 1\,\%$) typically gives errors of the order of 10 % on the periodic component (i.e., $\sigma_N \leq 20\,\mathrm{dB}$).

B. Theoretical analysis

Yegnanarayana et al. assume that $D_w(k) = C_w(k)$ for $k \in B_d$ ([14], p.5 col. 1, para 4), which implies that the periodic spectrum $G_w(k)$ is precisely zero for those frequency bins. Using the argument of compactness that they employ in Eqs. 16 and 17 (col. 2, bottom), it can be seen that the spectrum is zero at all frequencies: $G_w(k) = 0\ \forall\, k$. Yet, the authors remark that "the sidelobe effects of the windowing may produce significant values in the noise regions" ([14], p. 5). Therefore, provided that their argument is true, and that some part of the periodic component must reside in the aperiodic bins (as they remark), the convergent solution of the IIP must be the original:[6] $\lim_{m\to\infty} D_m(k) = C_w(k)$. In fact, the IIP, which is based on Parseval's theorem, is a standard signal reconstruction technique [56]. However, Eqs. 12, 13 and 14 should not be strict inequalities, since $D_m(k)$ equals $D_w(k)$ at convergence.[7] So, while the expressions guarantee that the error does not increase, they alone cannot guarantee that they converge on a unique solution, a point noted in [56].

[6] Otherwise, the convergence point would not be reached, no interpolation would take place and the solution would be (somewhat arbitrarily) determined by the initial assignment of bins.

[7] In the Papoulis-Gerchberg extrapolation technique from which this method is derived [57], the convergence region is explicitly excluded from the proof for this very reason.

C. Simulations

In their trials [14], [55], Yegnanarayana et al. evaluated the PAPD (Hamming window, $N = 512$ or 1024 points, sampling rate $f_s = 8\,\text{kHz}$) by the measured HNR and a perceptual spectral distance. We ran simulations of the PAPD using their parameters (Hamming, 512-point DFT, 8 kHz) on a mid-vowel section of the first vowel in /pɑzɑ/ recorded by an adult male speaker of British English RP, which was 6:1 downsampled to allow direct comparison with [14]. The signal was LPC pre-emphasized (10 pole, autocorrelation) and 255 points used for the analysis. At each iteration m of the interpolation, the signal power in the periodic and aperiodic estimates, $\langle d_m^2 \rangle$ and $\langle g_m^2 \rangle$, were calculated and plotted.

The results showed that the aperiodic estimate d_m began to approach convergence after about 1000 iterations, rather than after 20 as proposed [14]. Moreover, the solution upon which it appeared to converge was the original excitation signal, $c_w(n)$, suggesting that the algorithm, rather than decomposing the speech into periodic and aperiodic parts, actually reconstructed the original signal using half of the Fourier coefficients. Repeating the tests at other parts of the utterance revealed the same behavior. A second series of simulations was performed with signals synthesized from a pulse train plus GWN at HNRs ranging from -20 to $\infty\,\text{dB}$. Being spectrally flat, these signals required no LPC processing. Although convergence appeared to need a greater number of iterations, the results were similar: the IIP reconstructed the original signal, rather than achieving a stable decomposition.

Figure 10 shows the effect of IIP on the decomposed components (top) and the PAPD performance (bottom), for a pulse train in GWN. Again, the parameters specified in [14] were used (Hamming, 512, 8 kHz). As with the other examples, the aperiodic estimate converged to the original signal, the periodic estimate to zero, and the error to the original periodic component. The performance, despite showing a marginal improvement initially in this case, suffered severe degradation as the interpolation process was iterated, falling by 4 dB (η_v from 0 dB to -4 dB, η_u from 5 dB to 1 dB). By comparison, the PSHF achieved performances of $\eta_v = 1\,\text{dB}$ and $\eta_u = 7\,\text{dB}$ on the same example. Reconstruction of the original signal from the initial aperiodic estimate $d_0(n)$ was consistently observed for all trials over a wide range of noise levels, with different f_0 values, DFT sizes and window functions. The initial conditions and the rate of convergence varied depending on the original signal's real cepstrum, which was governed by the choice of window and the details of the noise, but the asymptotic behavior appeared in every case. Thus,

because of the theoretical aspects that were overlooked, and the low number of iterations used, the PAPD algorithm appears to yield a reasonable decomposition.

D. Comparison

For the sake of a direct comparison of the PAPD algorithm against our PSHF, we present the results of decomposing synthetic signals using each technique. The signals were generated as described in [55]: the authors of [55] provided us with their decomposition for comparison, while the PSHF was applied by us. The four signals were chosen to be representative of the wider tests with noise, jitter, shimmer and an alternative f_0. They are listed in Table III along with the corresponding performance results. The prescribed (initial) and measured (decomposed) HNRs are also given.

For these test signals, the performance results of the PSHF are better than those of the PAPD in all cases by at least 3 dB, and the HNRs estimated from the signals are more accurate in all but one case. (To validate our own implementation of the PAPD algorithm used earlier in this appendix, we compared our output signals with those supplied by d'Alessandro et al. and obtained very similar results.)

E. Discussion

Our simulations, running over 10^5 iterations, support our theoretical argument that convergence is achieved when the original signal has been reconstructed (Section A-B). This result was expected since similar algorithms have successfully been applied to incomplete spectra as a solution to signal reconstruction problems [56], [58], [59]. The iterative algorithm is, however, a cumbersome, computationally-intensive method of signal reconstruction.[8] The initial estimate of the aperiodic part is based on the assumption that $D(k) = A_w(k)$ for $k \in B_d$, but this does not account for the effects of windowing, despite the authors' earlier remark that it is an important issue. The PAPD crucially depends on the interplay between the spectral leakage of the rectangular window and the original Hamming window to determine the rate of convergence. Therefore, the amount of energy in the interpolated bins ($k \notin B_d$) depends on the number of iterations, on the way the time-compactness criterion is enforced (i.e., by rectangular window), on

[8]For this purpose, it could probably be replaced by a single-step calculation, like other such methods [56], involving the Hilbert transform or convolution with the complex spectrum of the rectangular window $\xi(n)$ that enforces the time-compactness criterion, for instance.

the HNR (owing to side-lobe leakage of the periodic part into the initial estimate $D_0(k) = A_w(k)$, for $k \in B_d$), and on the details of the aperiodic spectrum in the initial estimate. It does not depend on the amount of aperiodic energy any more than on the periodic energy. It was probably chance that the interpolated aperiodic energies approximated the expected HNR values. Indeed, the discrepancies observed in Fig. 3 of ([55], Section IIIA, p.18) can be explained by this and by the decision to iterate only twenty times.

Although it flattens the formant peaks, pre-whitening the speech signal by LPC is unable to help when the sources have differing spectral tilts, and when additional zeros are present in the noise spectrum, which are both typical features of voiced consonants. Moreover, in frequency regions of low HNR, voicing makes a negligible contribution and yet the PAPD allocates, on average, half of the excitation energy to the initial periodic estimate. While harmonics and noise are spectrally indistinct, the PSHF allocates one fourth of the excitation energy to the harmonic estimate. The authors state "that the decomposition algorithm is able to separate aspiration noises and the periodic noise in the voice source" ([14], p. 9). However, the low sampling rate used (8 kHz) means that much of the turbulence noise was missed. These factors hindered the PAPD's ability to uncover new features in the decomposed speech.

In light of the above mathematical argument and the simulation results, we conclude that the PAPD ultimately converges, if the objective is a decomposition, to the wrong solution. Nonetheless, while there are some critical flaws and several shortcomings in the PAPD algorithm [14], it may still be applicable as an analysis tool, and their use of synthetic signals and choice of variables offer a comprehensive methodology for testing decomposition algorithms, which we have largely followed.

References

[1] K. N. Stevens, "Models for the production and acoustics of stop consonants," *Speech Comm.*, vol. 13, pp. 367–375, 1993.

[2] C. Scully and S. J. Mair, "Relationships between different descriptive frameworks for plosive features of voicing and aspiration," *Levels in Speech Communication: Relations and Interaction,* eds. J. Schoentgen, J. M. Ramlot, C. Sorin, H. Meloni and J. Mariani, Elsevier Science B. V., pp. 51–62, 1995.

[3] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, Netherlands, 1960.

[4] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, Berlin, 2nd edition, 1972.

[5] K. N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1998.

[6] H. Herzel, "Bifurcations and chaos in voice signals," *Appl. Mech. Rev.*, vol. 46, no. 7, pp. 399–413, 1993.

[7] E. Yumoto, W. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.*, vol. 71, no. 6, pp. 1544–1550, 1982.

[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.*, vol. 84, no. 4, pp. 1292–1301, 1988.

[9] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *J. Speech & Hearing Res.*, vol. 36, pp. 254–266, 1993.

[10] S. N. Awan and M. L. Frenkel, "Improvements in estimating the harmonics-to-noise ratio of the voice," *J. Voice*, vol. 8, no. 3, pp. 255–262, 1994.

[11] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio — a new measure for describing pathological voices," *Acta Acustica*, vol. 81, pp. 700–706, 1995.

[12] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *J. Acoust. Soc. Am.*, vol. 102, no. 1, pp. 537–543, 1997.

[13] Y. Qi, R. E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *J. Acoust. Soc. Am. Lett.*, vol. 105, no. 4, pp. 2532–2535, 1999.

[14] B. Yegnanarayana, C. R. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. SAP*, vol. 6, no. 1, pp. 1–11, 1998.

[15] P. J. Murphy, "Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2866–2881, 1999.

[16] F. M. Silva and L. B. Almeida, "Speech separation by means of stationary least-squares harmonic estimation," *Proc. IEEE-ICASSP*, vol. 2, pp. 809–812, 1990.

[17] J. T. Graf and N. Hubing, "Dynamic time warping comb filter for the enhancement of speech degraded by white Gaussian noise," *Proc. IEEE-ICASSP*, vol. 2, pp. 339–342, 1993.

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

[19] R. I. Damper, J. R. Thorpe, and C. H. Shadle, "Separation of speech from simultaneous talkers," *Proc. 13th Int. Cong. Phon. Sci., Stockholm, Sweden*, vol. 3, pp. 282–285, 1995.

[20] C. D. Yoo and J. S. Lim, "Speech enhancement based on the generalised dual excitation model with adaptive analysis window," *Proc. IEEE-ICASSP*, pp. 832–835, 1995.

[21] B. T. Logan and A. J. Robinson, "Enhancement and recognition of noisy speech within an autoregressive hidden Markov model framework using noise estimates from the noisy signal," *Proc. IEEE-ICASSP*, 1997.

[22] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," *Proc. IEEE-ICASSP*, vol. 93, no. 2, pp. 550–553, 1993.

[23] Y. Stylianou, *Harmonic plus Noise Models for Speech, Combined with Statistical Methods for Speech and Speaker Modification*, Ph.D. thesis, Signals Dept., ENST-Telecom, Paris, 1995, ftp://ftp.research.att.com/dist/stylianou/thesis.ps.gz.

[24] G. Richard and C. R. d'Alessandro, "Modification of the aperiodic components of speech signals for synthesis," *Progress in Speech Synthesis,* eds. J. P. H. van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg, Springer-Verlag, Berlin, pp. 41–56, 1997.

[25] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition," *Comp. Mus. J.*, vol. 14, no. 4, pp. 12–24, 1990.

[26] P. Cook, "Noise and aperiodicity in the glottal source: A study of singer voices," *Proc. 12th Int. Cong. on Phon. Sci., Aix-en-Provence, France*, pp. 166–170, 1991.

[27] M. Feder, "Parameter estimation and extraction of helicopter signals observed with a wide-band interference," *IEEE Trans. on Sig. Proc.*, vol. 41, no. 1, pp. 232–244, 1993.

[28] D. L. Donoho, "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data," *Proc. Sym. in App. Math.*, vol. 47, pp. 173–205, 1993.

[29] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," *Proc. IEEE-ICASSP*, pp. 251–253, 1976.

[30] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. ASSP*, vol. 26, no. 4, pp. 354–358, 1978.

[31] E. N. Pinson, "Pitch-synchronous time-domain estimation of formant frequencies and bandwidths," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1264–1273, 1963.

[32] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, no. 4, pp. 911–918, 1976.

[33] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. ASSP*, vol. 36, no. 8, pp. 1223–1235, 1988.

[34] V. Darsinos, C. R. d'Alessandro, and B. Yegnanarayana, "Evaluation of a periodic/aperiodic speech decomposition algorithm," *Proc. Eurospeech '95, Madrid, Spain*, pp. 393–396, 1995.

[35] C. R. d'Alessandro, B. Yegnanarayana, and V. Darsinos, "Decomposition of speech signals into deterministic and stochastic components," *Proc. IEEE-ICASSP*, pp. 760–763, 1995.

[36] D. C. Rife and R. R. Boorstyn, "Single-tone parameter estimation from discrete-time observations," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 591–598, 1974.

[37] M. B. Priestley, *Spectral Analysis and Time Series*, Probability and Mathematical Statistics. Academic Press, London, UK, 1981.

[38] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Lecture Notes in Statistics. Springer-Verlag, Berlin, FRG, 1988.

[39] D. C. Rife and R. R. Boorstyn, "Multiple tone parameter estimation from discrete-time observations," *Bell Sys. Tech. J.*, pp. 1389–1410, 1976.

[40] G. M. Jenkins and D. G. Watts, *Spectral Analysis and its applications*, Time Series Analysis. Holden-Day, San Francisco, CA, 1968.

[41] 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, vol. 1, pp. 263–264, 1998.

[42] D. J. Hermes, "Measurement of pitch by sub-harmonic summation," *J. Acoust. Soc. Am.*, vol. 83, no. 1, pp. 257–264, 1988.

[43] W. Hess, *Pitch determination of speech signals: algorithms and devices*, Springer-Verlag, Berlin, 1983.

[44] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, pp. 293–309, 1967.

[45] J. P. Dworkin and R. J. Meleca, *Vocal Pathologies*, Singular Pub., San Diego, CA, 1997.

[46] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *J. Acoust. Soc. Am.*, vol. 103, no. 5 Pt. I, pp. 2649–2658, 1998.

[47] Y Horii, "Fundamental frequency perturbations observed in sustained phonation," *J. Speech & Hearing Res.*, vol. 22, pp. 5–19, 1979.

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

[49] P. H. Dejonckere and J. Lebacq, "Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology," *J. ORL*, vol. 58, no. 6, pp. 326–332, 1996.

[50] D. M. Howard and A. J. Fourcin, "Instantaneous voice period measurement for cochlear stimulation," *Electronics Lett.*, vol. 19, no. 19, pp. 776–778, 1983.

[51] P. J. B. Jackson and C. H. Shadle, "Frication noise modulated by voicing, as revealed by pitch-scaled decomposition," *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1421–1434, 2000.

[52] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[53] C. R. d'Alessandro, "Time-frequency speech transformation based on an elementary waveform representation," *Speech Comm.*, vol. 9, no. 5/6, pp. 419–431, 1990.

[54] P. J. B. Jackson, *Nephthys project*, Sch. Electron. and Elec. Eng., Univ. Birmingham, UK, 2000, http://web.bham.ac.uk/p.jackson/nephthys/.

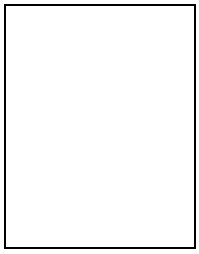
[55] C. R. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Trans. SAP*, vol. 6, no. 1, pp. 12–23, 1998.

[56] M. H. Hayes, J. S. Lim, and A. V. Oppenhiem, "Signal reconstruction from phase or magnitude," *IEEE Trans. ASSP*, vol. 28, no. 6, pp. 672–680, 1980.

[57] A. Papoulis, *Signal Analysis*, McGraw-Hill, New York, NY, 1984.

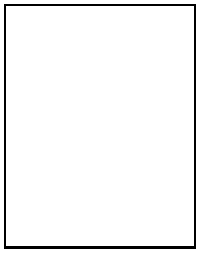
[58] J. C. Anderson, "Complex signal reconstruction from time-frequency magnitude," *IEEE Trans. ASSP*, vol. 6, pp. 297–300, 1994.

[59] A. M. Taratorin and S. Sideman, "Signal reconstruction from noisy-phase and magnitude data," *Applied Optics*, vol. 33, no. 23, pp. 5415–5425, 1994.



Philip Jackson (S'97–M'01) received the BA Hons degree in engineering in 1993 from the University of Cambridge, UK, and the PhD in electronics in 2000 from the University of Southampton, UK. He worked from 1993 to 1996 as a research and flight-test engineer on active noise control systems for aircraft with Noise Cancellation Technologies, later Ultra Electronics. From 2000 to the present he has been a Research Fellow at the School of Electronic and Electrical Engineering, University of Birmingham, working on articulatory approaches to automatic speech recognition. His research interests also include vocal-tract modeling, turbulence-noise sources in speech production, and speech analysis.

Dr. Jackson is a member of the Acoustical Society of America and the International Speech Communication Association, and an associate member of the Institution of Electrical Engineers (UK).



Christine Shadle (S'75–M'85) received an AB in music and MS in electrical engineering in 1976 from Stanford University, and the PhD in electrical engineering in 1985 from MIT. From 1976 to 1979 she worked at Bell Telephone Laboratories, Murray Hill, on LPC analysis and re-synthesis. From 1985 to 1987 she was a Hunt Fellow (of the ASA) and a NATO Fellow at the ISVR, University of Southampton, UK, and the Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden. In 1987 she became a Lecturer, and in 1999 a Senior Lecturer, at the Department of Electronics and Computer Science, University of Southampton, UK. In 1995–1996 she was a Guest Researcher at the Human Information Processing Laboratory, ATR, Kyoto, Japan. Her research interests include acoustics and aeroacoustics of speech production, and vocal tract imaging.

Dr. Shadle is a member of the Acoustical Society of America, the International Speech Communication Association, and Sigma Xi. She has been an elected member of the Acoustical Society of America Technical Committee on Speech Communication, and is a member of the Permanent Council of the International Conference on Voice Physiology and Biomechanics.

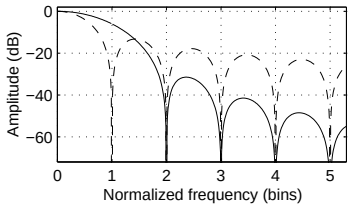


Fig. 1. Smearing effect on the spectral envelope of rectangular (dashed) and Hanning (solid) windows. [recommended width = 1 col. (11pc)]

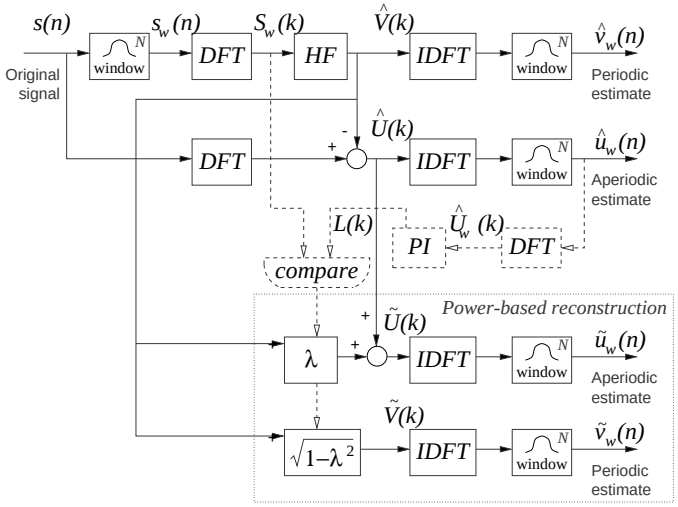


Fig. 2. The pitch-scaled harmonic filter (PSHF) algorithm. The top half provides one periodic/aperiodic pair of output signals for time-series analysis, using the harmonic filter (HF), while the bottom half gives a pair for power spectral analysis, after performing the power interpolation (PI). [recommended width = 1 col. (21pc)]

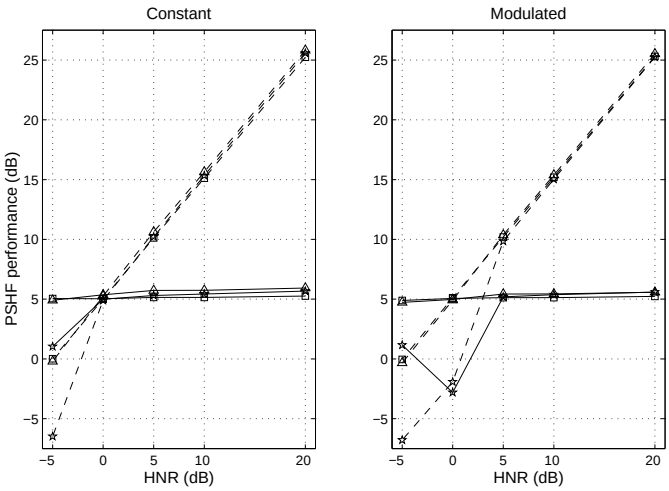


Fig. 3. Aperiodic η_u (dashed) and periodic η_v (solid) performance of the PSHF on synthetic speech signals versus HNR; with constant (left) and modulated (right) noise. Each graph shows results for three values of f_0: 120 Hz ($\triangle$), 130.8 Hz (star), 200 Hz (box). No jitter or shimmer. See text for values at $\sigma_N = \infty$ dB. [recommended width = 1 col. (21pc)]

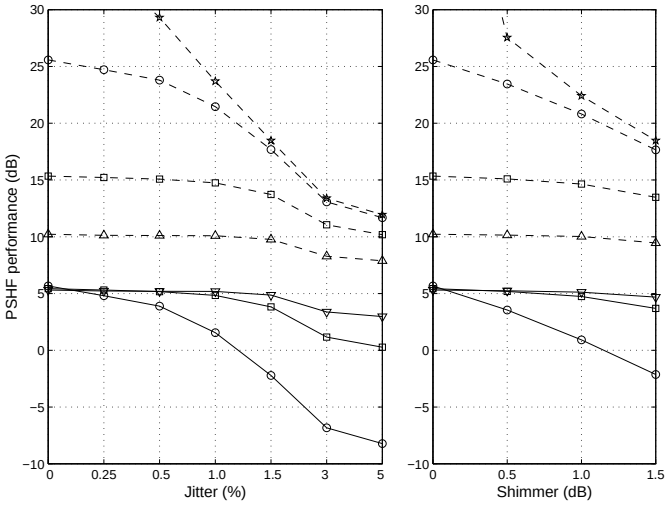


Fig. 4. Aperiodic η_u (dashed) and periodic η_v (solid) performance of the PSHF on synthetic speech signals, perturbed with either jitter (left) or shimmer (right). For both, the HNRs are: ∞ dB (star), 20 dB ($\circ$), 10 dB (box), or 5 dB ($\triangle$). [recommended width = 1 col. (21pc)]

TABLE I

PERIODIC AND APERIODIC PERFORMANCE OF THE PSHF VERSUS JITTER σ_T, SHIMMER σ_A AND HNR σ_N. ENTRIES ARE $(\eta_v\ \eta_u)$ IN DB. [RECOMMENDED WIDTH = 1 COL. (21PC)]

σ_T	σ_A	Initial HNR (dB)							
%	dB	∞		20		10		5	
0	0	–	54	6	26	5	15	5	10
	1	–	22	1	21	5	15	5	10
0.5	0	–	29	4	24	5	15	5	10
	1	–	19	−2	18	4	14	5	10
3	0	–	13	−7	13	1	11	3	8
	1	–	14	−6	13	2	11	4	9

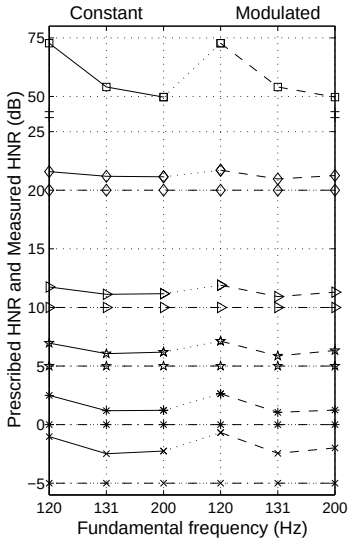


Fig. 5. Measured HNR for constant (solid) and modulated (dashed) noise versus f_0, shown for the prescribed values (dash-dot, from bottom): −5 dB (×), 0 dB (∗), 5 dB (star), 10 dB (△), 20 dB (♢), ∞ dB (box, separate scale). No jitter or shimmer. [recommended width = 1 col. (11pc)]

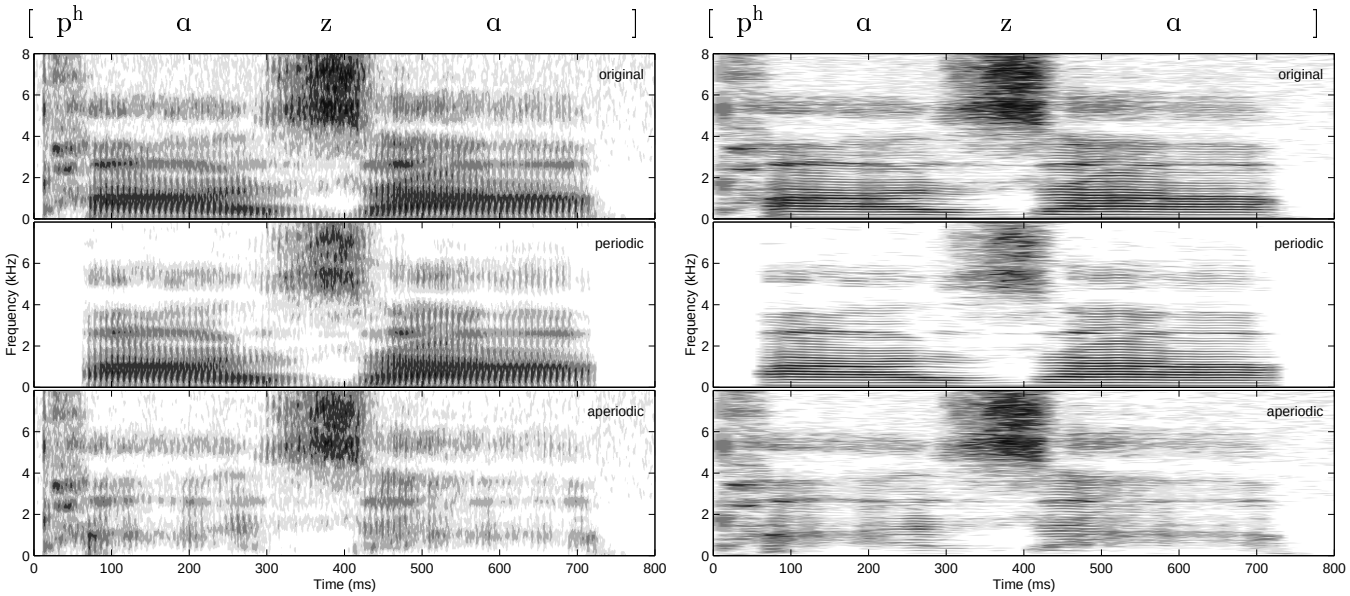


Fig. 6. Wide-band (left) and narrow-band (right) spectrograms (5 ms and 43 ms respectively, Hanning window, ×4 zero-padded, fixed grey-scale) of the utterance [pʰɑzɑ] by an adult male speaker (PJ): from (top) the original signal $s(n)$, (middle) the periodic estimates $\hat{v}(n)/\tilde{v}(n)$, and (bottom) the aperiodic estimates $\hat{u}(n)/\tilde{u}(n)$. [recommended width = 2 col. (7.1875in)]

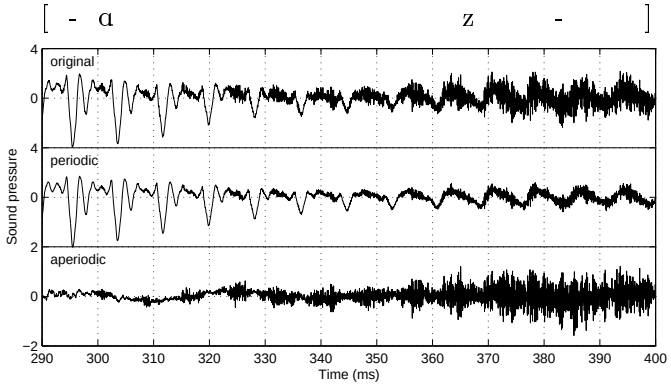


Fig. 7. Time series of the original signal $s(n)$ (top) from the vowel-fricative transition [-ɑz-] by an adult male speaker (PJ), the periodic component $\hat{v}(n)$ (middle) and the aperiodic component $\hat{u}(n)$ (bottom, note double amplitude scale). [recommended width = 1 col. (21pc)]

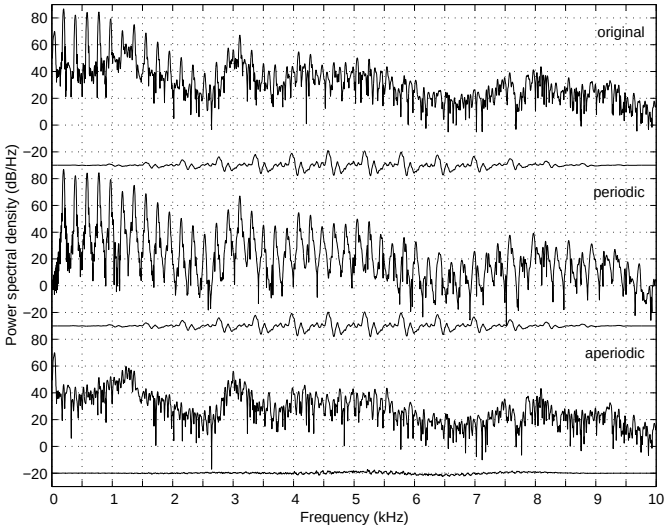


Fig. 8. Power spectra (85 ms, Hanning window, ×4 zero-padded) computed from the original signal $s(n)$ (top) from the vowel [ɑː] by an adult female speaker (SB), the periodic estimate $\tilde{v}(n)$ (middle) and the aperiodic estimate $\tilde{u}(n)$ (bottom), whose time series are inset underneath each graph (aperiodic signal double scale). [recommended width = 1 col. (21pc)]

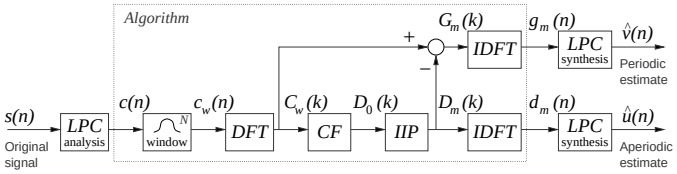


Fig. 9. The periodic-aperiodic decomposition (PAPD) algorithm, whose core comprises a cepstral filter (CF) and the iterative interpolation process (IIP). [recommended width = 1 col. (21pc)]

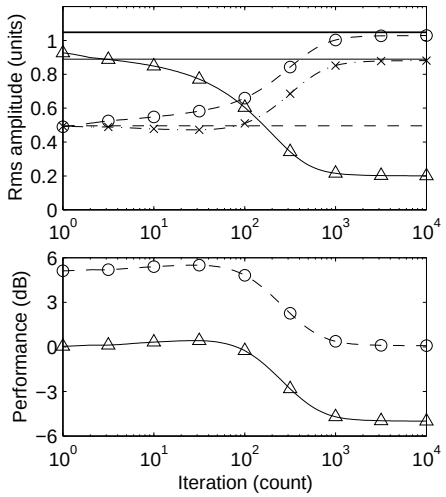


Fig. 10. Effect of the PAPD's iterative process. Top: log-linear plot of the root mean square amplitude of the periodic estimate g_m ($\triangle$, solid), the aperiodic estimate d_m ($\circ$, dashed) and the error ($\times$, dash-dot) versus iteration count, for a pulse train ($f_0 = 120\,\mathrm{Hz}$) in Gaussian white noise (HNR = 5 dB). The horizontal lines indicate the original signal (thick, solid) with its components (thin): periodic (solid) and aperiodic (dashed). Bottom: the periodic ($\triangle$, solid) and aperiodic ($\circ$, dashed) performance in dB. [recommended width = 1 col. (16pc)]

TABLE II

KEY TO SYMBOLS USED HERE AND IN [14]. [RECOMMENDED WIDTH = 1 COL. (21PC)]

Here	[14]	Description
n	n	point in time
k	k	point in frequency
N	N	DFT length
m	m	iteration number
B_d	F_r	selected aperiodic bins
$c(n)$	$e(n)$	excitation signal
$C_w(k)$	$E(k)$	excitation spectrum
$d(n)$	$r(n)$	true aperiodic excitation signal
$D_w(k)$	$R(k)$	true aperiodic part's spectrum
$G_w(k)$	$P(k)$	true periodic part's spectrum
$d_0(n)$	$r_0(n)$	initial estimate of aperiodic signal
$D_0(k)$	$R_0(k)$	initial estimate of aperiodic spectrum
$'$	$\hat{}$	time-compacted version
$d_m(n)$	$r_m(n)$	mth estimate of aperiodic signal
$D_m(k)$	$R_m(k)$	mth estimate of aperiodic spectrum
$g_m(n)$	$p_m(n)$	mth estimate of periodic signal (p. 2)
$g_m(n)$	$g_m(n)$	mth estimate of periodic signal (p. 5)
$\delta(n)$	$l(n)$	hypothetical signal
$\Delta(k)$	$L(k)$	hypothetical spectrum
$\gamma(n)$	$h(n)$	hypothetical periodic signal
$\Gamma(k)$	$H(k)$	hypothetical periodic spectrum

TABLE III

COMPARATIVE RESULTS, WHERE THE BURST DURATION RATIO (BDR) IS THE PROPORTION OF EACH PITCH PERIOD FOR WHICH THERE IS NOISE. [RECOMMENDED WIDTH = 2 COL. (7.1875IN)]

Settings					(dB)	PAPD			PSHF		
f_0	J	S	Nom.	BDR	Init.	Decomp.	Performance		Decomp.	Performance	
Hz	%	dB	HNR	%	HNR	HNR	η_v	η_u	HNR	η_v	η_u
120	0	0	10	60	11.7	15.2	1.1	12.7	14.5	4.4	16.1
120	1	0	∞	–	∞	32.0	$-\infty$	20.0	28.9	$-\infty$	27.5
120	0	0.5	∞	–	∞	17.4	$-\infty$	15.2	26.7	$-\infty$	23.0
200	0	0	5	100	5.47	8.5	2.8	8.2	7.9	5.8	11.2