Frication noise modulated by voicing, as revealed by pitch-scaled decomposition.

Philip J.B. Jackson* and Christine H. Shadle†

*School of Electronics and Electrical Engineering, University of Birmingham, Birmingham B15 2TT, UK. [p.jackson@bham.ac.uk][0]

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

Internet: http://www.isis.ecs.soton.ac.uk/research/projects/nephthys/

Short title: [Modulation of voiced fricatives]
Received:

Abstract

A decomposition algorithm that uses a pitch-scaled harmonic filter was evaluated using synthetic signals and applied to mixed-source speech, spoken by three subjects, to separate the voiced and unvoiced parts. Pulsing of the noise component was observed in voiced frication, which was analyzed by complex demodulation of the signal envelope. The timing of the pulsation, represented by the phase of the anharmonic modulation coefficient, showed a step change during a vowel-fricative transition corresponding to the change in location of the sound source within the vocal tract. Analysis of fricatives /β, v, ð, z, ʒ, ɣ, ʕ/ demonstrated a relationship between steady-state phase and place, and f_0 glides confirmed that the main cause was a place-dependent delay.
PACS numbers: 43.70.Bk, 43.72.Ar

1 Introduction

The production of voiced fricatives involves two predominant sources of sound exciting the vocal-tract resonances: the phonation source, produced by vocal-fold oscillation, and the noise source, produced downstream of a supraglottal constriction. Thus, if we wish to determine source characteristics from the speech signal, the analysis problem is more complicated than for single-source speech sounds and, as some authors have noted, the two sources are not entirely independent. In particular, the voicing source appears to modulate the noise source (Fant 1960; Flanagan 1972). Others have found that modulating the aspiration source during a vowel-to-voiced fricative transition leads to better-quality synthesis (Klatt and Klatt 1990; Scully 1990; Scully et al. 1992). While such interaction of sources inevitably complicates the model used for synthesis, and the analysis problem, it may also be the key to a more accurate model of the production mechanism itself. Closer study of the source interaction could lead directly to better quality synthesis of voiced fricatives and, potentially, of other mixed-source signals, such as breathy vowels.

In simple models of voiced fricatives, the voicing and frication sources are inserted into the system and the output is formed from the sum of their individual contributions: voicing as a volume velocity source at the glottis; frication as a pressure source at the supraglottal constriction. Although Fant (1960) noted that source-source interaction occurred as "periodic and synchronous" modulation of the frication source by phonation, Flanagan's electrical analog model was one of the first to incorporate modulation of the fricative source amplitude (Flanagan and Cherry 1969). Band-passed Gaussian noise (0.5–4 kHz) was multiplied by the square of the volume velocity at the constriction exit U_n, which included the d.c. component, to give the pressure (voltage) source P_n in series with a variable source resistance R_n. Sondhi and Schroeter (1987) employed a similar model for a practical implementation of an aspiration source at the glottis, gated by a threshold Reynolds number; for frication they placed a volume velocity source P_n/R_n one section (0.5 cm) downstream of the constriction exit (or at the lips for /f, v, θ, ð/), because of poor subjective results with pressure sources.

[0]The paper was written while the first author was at Dept. Electronics & Comp. Sci., Univ. of Southampton.

Scully (1990; Scully et al. 1992) based her source generation on Stevens' (1971) result from static experiments: the strength of the pressure source $p_s \propto \Delta P^{\frac{3}{2}}$, where ΔP is the pressure across the constriction. This source, depending on slowly-varying articulatory and aerodynamic parameters, was applied equally to aspiration and frication sources. Since ΔP across the supraglottal constriction is lower for voiced than voiceless fricatives, this equation partially accounts for the weaker frication source. These parameters do not encode any modulation, or allow for the flow separation lag in jet formation (Pelorson et al. 1997). However, motivated by the results of perceptual tests, the aspiration source was modulated using the rapidly-varying glottal area. Klatt, treating aspiration and frication identically, modulated the noise source with a square wave (50 % burst duration) that was switched on during voicing, remarking that it is "not necessary to vary the degree of amplitude modulation ..., but only to ensure that it is present" (Klatt 1980). In an analysis-by-synthesis procedure, Narayanan and Alwan (1996) used a combination of pressure (dipole) and volume velocity (monopole) sources to match measured fricative spectra, and concluded that the monopoles should be placed at the constriction exit and the dipoles at one or more obstacles: at the lips for /f, v, θ, ð/, at the teeth for /s, z/ and at the teeth and vocal-tract wall for /ʃ, ʒ/.

None of the above models considers any non-acoustic fluid motion, yet in a flow duct experiment (Coker et al. 1996), the arrival time of a pulse of radiated noise, depending strongly on the constriction-obstacle distance, suggested a convection velocity of less than half the flow velocity at the jet exit (8 m/s). In his recent PhD thesis, Sinder (1999) presents a model for fricative production that is based on aeroacoustic theory. Once the necessary flow-separation conditions have been met, vortices are shed, which convect along the tract, generating sound as they go, particularly when encountering an obstacle. Therefore, we want to consider both acoustic and aerodynamic mechanisms.

We have previously described an algorithm, the pitch-scaled harmonic filter (PSHF), that decomposes speech into harmonic and anharmonic signals (Jackson and Shadle 1998), which are estimates of the voiced and unvoiced components respectively. The PSHF was developed from a measure of harmonics-to-noise ratio (HNR, Muta et al. 1988) to provide full reconstruction of harmonic (voiced) and anharmonic (unvoiced) time series, on which subsequent analyses can be performed independently. This method is especially suited to acoustic analysis of sustained sounds with regular voicing (i.e. low values of jitter and shimmer), because of the underlying harmonic model of the voiced part, which is based on optimal (maximum likelihood) estimation. Other than the choice of the number of pitch periods (which is typical for adaptive filtering techniques), the PSHF is without any arbitrary features for heuristic adjustments, such as cut-off frequency (Laroche et al. 1993) and number of cepstral coefficients (Qi and Hillman 1997; Yegnanarayana et al. 1998), and does not suffer the bias, harmonic interference and variable performance problems of asynchronous harmonic techniques (Hardwick et al. 1993; Laroche et al. 1993; Qi and Hillman 1997; Serra and Smith 1990; Silva and Almeida 1990; Yegnanarayana et al. 1998).

In this paper, we employ the PSHF to study the interaction between sources in voiced fricatives, to arrive at better source models, and to obtain clues to the production mechanism that governs the interaction. Section 2 describes the PSHF method and tests of it using synthetic signals. Section 3 describes the recording method, subjects, and corpus, and presents preliminary results of the decomposition. Section 4 presents further analysis by considering the modulation of the aperiodic component in voiced fricatives, for which results are given in Section 5. These results are discussed in light of possible aeroacoustic mechanisms in Section 6, and Section 7 concludes.

2 Decomposition method

2.1 Pitch-scaled harmonic filter

The pitch-scaled harmonic filter (PSHF) was designed to separate harmonic and anharmonic components, $v(n)$ and $u(n)$, of a recorded speech signal $s(n)$. It assumes that these components will be representative of the acoustic consequences of the voiced and unvoiced sound sources respectively, i.e. the vocal-tract filtered excitations. A detailed description of the PSHF, including pitch estimation, windowing sequence and algorithm, can be found elsewhere (Muta et al. 1988; Jackson and Shadle 1998).Here we present a schematic summary of the central process, illustrating it with some spectra, followed by a description of tests using synthetic speech-like signals.

In the PSHF, the original speech signal $s(n)$ is decomposed primarily into the harmonic and anharmonic estimates, $\hat{v}(n)$ and $\hat{u}(n)$, respectively. Further harmonic and anharmonic estimates, $\tilde{v}(n)$ and $\tilde{u}(n)$, are

computed based on a power interpolation (PI) of the anharmonic spectrum, which improves the spectral composition of the signals when considering features over a time-frame longer than two pitch periods. Figure 1 describes the PSHF algorithm, which takes a 4-pitch period windowed section of the signal $s_w(n)$, transforms it into $S_w(k)$ by discrete Fourier transform (DFT), and decomposes it in the frequency domain by a harmonic filter (HF).[1] The output signals are then constructed by transforming the spectra $\hat{V}(k)$ and $\hat{U}(k)$ back into the time domain (by inverse DFT or IDFT) and windowing.

Figure 2 illustrates the operation of the harmonic filter using a mid-vowel recording of [ɑ] by an adult male (from example #1 by PJ, see Section 3 for details). It shows the original spectrum $S_w(k)$ after windowing, the spectrum of the harmonic estimate $\hat{V}_w(k)$, and the remainder $\hat{U}_w(k)$, the anharmonic spectrum. The essence of this technique is that, by scaling the window size to exactly four pitch periods $N = 4T_0$, the voiced (quasi-periodic) part is concentrated into every fourth bin of the spectrum. The pitch estimation process finds the value of T_0 that optimizes the concentration. Thus, a harmonic comb filter that passes these harmonic bins (and doubles them) yields an estimate of the voiced component $\hat{V}(k)$ which, after applying an IDFT, results in a periodic signal of length $4T_0$. Finally, the envelope of the estimate $\hat{v}_w(n)$ is matched to that of the input signal $s_w(n)$ by applying the same window function. The spectral consequences can be seen in Figure 2, which shows how, for each harmonic, the Fourier coefficient (middle) maintains the same value as that of the orignal spectrum (top), but has spread to the adjacent bins (at $-6\,\mathrm{dB}$). The residue is the anharmonic component $\hat{u}_w(n)$, whose spectrum (Fig. 2, bottom) accordingly contains gaps at the harmonics. For a periodic signal in Gaussian white noise, the harmonic Fourier coefficients provide the optimal (maximum likelihood) estimate of the signal (Rife and Boorstyn 1974; Bretthorst 1988) and the residue is thus the best estimate of the noise. However, if one is interested in the anharmonic power spectrum, particularly at a fine frequency resolution ($\leq f_0/2$), intuitively one would consider filling the gaps by some form of interpolation, assuming that the noise is the result of a stochastic process with a smoothly varying frequency response. So, the PI stage computes the mean power of the bins either side of each harmonic $L(k)$. Then, by comparing $L(k)$ with the original coefficients $S_w(k)$, the factor $\lambda(k)$ is used to share the power from the harmonic bins between the harmonic and anharmonic spectra, $\tilde{V}(k)$ and $\tilde{U}(k)$, giving new power-based estimates $\tilde{v}_w(n)$ and $\tilde{u}_w(n)$, respectively. An entire section of voiced speech can be processed by sliding the window along, and by overlapping and adding the outputs $\hat{v}_w$, $\hat{u}_w$, $\tilde{v}_w$ and $\tilde{u}_w$ to obtain complete signals $\hat{v}$, $\hat{u}$, $\tilde{v}$ and $\tilde{u}$.

2.2 Synthetic test signals

To use the PSHF for studying modulation of noise sources in detail, we need to ascertain the performance of the PSHF for such signals. Twelve speech-like test signals $s(n)$ were composed of a deterministic part $v(n)$ and a noise part $u(n)$:

$$s(n) = v(n) + u(n)\,, \tag{1}$$

at sampling rate $f_s = 48\,\mathrm{kHz}$. The deterministic part was synthesized by convolving a pulse train $g(n)$, which was periodic at $f_0 = 120.0\,\mathrm{Hz}$, with an appropriate impulse-response filter h:

$$v = g * h\,, \tag{2}$$

where $*$ denotes convolution. The filter h was built using the linear prediction coefficients (LPC, autocorrelation, 50-pole) obtained from the same adult male mid-vowel [ɑ] recording used in Fig. 2 (#1 by PJ). The noise signal was similarly created by convolving Gaussian white noise $d(n)$ (zero mean, unit variance) with the LPC filter. However, the noise was combined in two ways: with constant-variance noise, and with its amplitude modulated at the fundamental frequency, f_0:

$$u = \begin{cases} G\,(d * h) \\ G\,(d * h)\,\sqrt{\frac{2}{3}}\left[1 + \cos\left(\frac{2\pi f_0 n}{f_s} + \beta\right)\right]\,. \end{cases} \tag{3}$$

The modulation was set at phase $\beta \in \{0, \pi/4, \ldots, 7\pi/4\}$ in relation to the glottal excitation; the factor of $\sqrt{2/3}$ equalized the noise to give the same mean signal power. The gain G was adjusted to give harmonics-to-noise ratios (HNRs) at one of six specified levels: ∞, 20, 10, 5, 0 or $-5\,\mathrm{dB}$.[2]

Using the specified pitch as an initial estimate, the local minimum in the pitch-estimating cost function was found at a series of points throughout each test signal. For high HNRs, the estimated period was identical to the true T_0 but, as the noise level was increased, so did the deviation of the estimates. These values were given as the pitch input to the PSHF, which then processed each signal in the usual way: incrementing

the analysis frame, decomposing the signal and accumulating the outputs.[3] Thus, using the PSHF signal estimates $\hat{v}$ and $\hat{u}$, the changes in signal-to-error ratio (SER), η_v and η_u, were calculated as a measure of performance of the decomposition algorithm. For the harmonic component, the change in SER η_v is defined as the ratio of the initial noise to the residual error; conversely, the anharmonic performance η_u is the ratio of the deterministic part to the error. Both are expressed in decibels:

$$\eta_v = 10\log_{10}\left(\frac{\langle v^2\rangle/\langle e^2\rangle}{\langle v^2\rangle/\langle u^2\rangle}\right) = 10\log_{10}\left(\frac{\langle u^2\rangle}{\langle e^2\rangle}\right), \quad (4)$$

$$\eta_u = 10\log_{10}\left(\frac{\langle v^2\rangle}{\langle e^2\rangle}\right), \quad (5)$$

where the residual error is $e = (\hat{v} - v) = -(\hat{u} - u)$. Although these two expressions are clearly related by the HNR σ_N (i.e., $\eta_u = \sigma_N + \eta_v$), it is useful to describe the performance of both components separately.

Table 1 lists the harmonic (in parentheses) and anharmonic performance over the range of specified noise conditions. Except for the anharmonic performance at the -5 dB condition, all the performance values are positive, which implies that the quality of the separated component is better than the input signal, i.e., the remaining errors are always smaller than the original corruption from the interfering source. The anharmonic performance is strongly correlated with HNR, and is approximately 5 dB greater than the initial HNR, so that any residual errors in the extracted anharmonic signal are about half as large as the true noise component. Meanwhile, the harmonic component is cleaned up to a similar degree by the PSHF, which reduces the errors to about half of their original amplitude, on average. Note that the results of the constant-variance and modulated noise cases are almost identical for $\beta = 180°$, which implies that the performance is not significantly affected by the envelope of the noise. Tests at other phase settings produced similar results ± 0.2 dB. Overall, the results indicate the extent to which we can have confidence in the output signals that the PSHF produces.

Although there are transient errors for the first two pitch periods, as the tail of the first window ramps up towards its center, the decomposed components shown in Figure 3 soon approach the true components. Looking at the time series more closely, it is apparent that the modulation of the noise envelope is retained. Indeed, the error signal also exhibits some modulation, suggesting that the error is proportionally related to the noise, for a given mean HNR. The amplitude of the envelope of $\hat{u}$ is slightly reduced with respect to the input component u, but *its phase remains unaltered*. This finding, which is crucial to the results presented in this article, will be further justified in Section 4.4. These simulations, therefore, support the assertion that any modulation exhibited by the anharmonic component is not a processing artifact, but a property of the source component from which it is derived.

3 Application of PSHF

3.1 Recording details

A series of recordings was made by three adult subjects who had no known speech pathologies: two native British English speakers, one male (PJ) and one female (SB), and a Portuguese male (LJ). The speech corpus contained sustained fricatives (all subjects) and some additional items: sustained vowels /ɑ, i, u/ (PJ, LJ, SB), nonsense words /pʰɑFɑ/ (PJ, LJ, SB) and fricatives with f_0 glides /v, ð, z, ʒ/ (PJ). One subject, PJ, also recorded non-modal sustained vowels (viz. pressed, breathy and whispered). The sustained fricatives were placed in a vowel context /VF:/ and sustained for 5 s. The fricatives F:, given here in unvoiced-voiced pairs, were: /ɸ, β/ (bilabial), /f, v/ (labiodental), /θ, ð/ (dental), /s,z/ (alveolar), /ʃ, ʒ/ (palatoalveolar), /x, ɣ/ (velar), /ħ,ʕ/ (pharyngeal). None of the subjects was a trained phonetician, and none has all of the fricatives natively; the recordings nevertheless exhibit a range of place variation. The /pʰɑFɑ/ nonsense words were repeated to give 10 tokens using a single breath.

The sound pressure at 1 m was measured in a sound-treated room using a microphone (B & K 4165/4133), a pre-amplifier (B & K 2639) and amplifier (B & K 2636, 22 Hz–22 kHz band-pass, linear filter). An electroglottograph (EGG, Laryngograph PCLX) with large (adult) electrodes was used to measure the transglottal impedance. Both signals were recorded on DAT (Sony TCD-D7, $f_s = 48$ kHz), from which they were later digitally transfered to computer as 16-bit stereo data. A calibration tone and background noise were recorded with the microphone channel to give an absolute reference to pressure and to assess the measurement-error (noise) floor, respectively.

3.2 Decomposition of [pʰɑzɑ]

The utterance that we refer to as example #1 consisted of the nonsense word [pʰɑzɑ] spoken by subject PJ. To illustrate the effect of the PSHF, it was decomposed into its harmonic and anharmonic parts, as shown in Figure 4. The original signal (top) shows the initial burst (20 ms) followed by voice onset (70 ms), the first vowel (100–330 ms), the voiced fricative (320–420 ms) and the second vowel (420–720 ms). The harmonic component shows the voicing with reduced noise, as expected; the anharmonic component contains the burst transient and initial noise (20–70 ms), a small amount of noise during the vowels and a larger amount during the fricative.

The PSHF algorithm tended to provide the most faithful decomposition during steady spells of voicing, when the amplitude and fundamental frequency varied little. The presence of jitter, shimmer and abrupt changes causes perturbation errors, which can be seen in the anharmonic component (70–100 ms, 200 ms, 270 ms and 450 ms in Fig. 4).

Figure 5 comprises spectrograms of the signals, which contain the following features: vertical stripes at the glottal pulse instants, slowly-varying horizontal bands (the formant resonances), a generally mottled appearance of the anharmonic component indicative of a noisy signal, and the separation of voicing and frication during the voiced fricative. Note that, without the aid of any heuristic filtering, the majority of the high-frequency turbulence noise has been passed to the anharmonic component, while the low-frequency voiced part has been successfully allocated to the harmonic component. It is also possible to see vertical striations during the frication onset in the high-frequency turbulence noise, which become less noticeable mid-fricative.

Looking at the vowel-fricative transition in more detail (see Figure 6), we see the growth of the anharmonic component while the voicing dies down. Compared with the original signal, the harmonic component is much cleaner in appearance, and the regularity of the continuing vocal fold oscillation is obvious, even in the middle of the fricative (c. 380 ms), despite much weaker phonation. Although devoicing sometimes occurs in voiced fricatives, it is clear that that is not the case here. The anharmonic component $\hat{u}$, which is plotted with double the amplitude scale, is very small at the end of the vowel, commensurate with a typically high HNR for modal voice (+17 dB). The HNR drops dramatically by 20 dB, to about −3 dB, as $\hat{u}$ grows during the transition. We also see pulsing of the noise, which becomes less noticeable as the fricative develops; the noise initially comes in bursts with each glottal pulse, then disperses into continuous noise in the fully-developed fricative. Despite the inevitable degradation in PSHF performance, the disappearance of the modulation probably owes more to the decreased amplitude of phonation than to processing artifacts.

4 Modulation

4.1 Short-time power (STP)

By seeing how the envelopes of the harmonic and anharmonic signals vary over time, we can investigate not only the ratio of the two, the short-time HNR, but also their individual trajectories. Averaging over a frame comparable with a pitch period, we can see finer variations such as those of the anharmonic component caused by the modulation of the noise. The use of these derived measures is best demonstrated at the transition between a vowel and a mixed-source sound that has a strong anharmonic component, as illustrated by the vowel-fricative transition [-ɑz-] in Figure 6.

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

$$P_y(p) = \frac{\sum_{m=0}^{M-1} x^2(m) y^2(p+m-M/2)}{\sum_{m=0}^{M-1} x^2(m)}, \qquad (6)$$

using the smoothing window $x(m)$ of length M. Thus, P_v is the STP of the harmonic component and P_u that of the anharmonic component. The window x acts as a low-pass filter on the squared signals, whose roll-off frequency is governed by the window length M, which reduces the interference from higher harmonics. As such, periodic variations in STP are eliminated with the larger window, yet remain, albeit at a reduced amplitude (−6 dB), with the shorter window.[4] For each computation of the STP, we set M to a constant and used a Hanning window: $x(m) = \frac{1}{2}\left(1 - \cos\frac{2\pi m}{M}\right)$ for $m \in \{0, 1, \ldots, M-1\}$. In the present study, we were interested in features visible only at high time resolution (of order less than two pitch periods) so, although we were computing the (short-time) power from the signals to calculate P_v and P_u, $\hat{v}(n)$ and $\hat{u}(n)$ were used rather than the power-estimated $\tilde{v}(n)$ and $\tilde{u}(n)$, which are designed for narrow-band spectral analysis. In doing so, we were exploiting the PSHF's signal reconstruction in order to generate features from subsequent (asynchronous) analysis.

4.2 Observations of [ɑz:]

Speech example #2, the vowel-fricative transition [ɑz:] produced by subject PJ, was decomposed by the PSHF and the STPs were calculated. To observe short-term variations, the window length was set to the mean period, $M = \langle T_0 \rangle$; for medium-term variations over the length of the utterance, the window length was set to approximately four times the period, $M = 4\langle T_0 \rangle$.

The resultant STPs are plotted in dB in Figure 7. The difference between the harmonic and anharmonic medium-term STP trajectories (top) is the short-term HNR which, besides voice onset, shows a noticeable change at about 400 ms at the transition from vowel to fricative. Indeed, after voicing has peaked towards the beginning of the vowel (at about 160 ms), the harmonic amplitude dies away, reaching a maximum decay at the transition (circa 400 ms). After some overshoot and subsequent fluctuations it returns to a steady value (c. 700 ms). The anharmonic component grows during the development of the fricative (380–500 ms), undergoes a period of oscillation (500–660 ms) and finally settles down to a reasonably steady value. Note that the fluctuations of the two components at the start of the fricative are roughly equal and opposite. The initial period fluctuations at voice onset cause errors in the harmonic estimate, which get replicated, in negative, in the anharmonic estimate. Otherwise, the HNR is at least +10 dB in the vowel, rising to more than +20 dB at the steadiest point (around 200 ms). In the fricative, values range from −3 dB to +10 dB, settling to about +8 dB in the fully-established part. The short-term STP curves (Fig. 7, bottom), which were computed using the single-period smoothing window, exhibit the same general trends, but have an oscillating element superimposed, which is caused by the modulations in signal power within individual pitch periods.

4.3 Pitch-scaled demodulation

In order to quantify the oscillations in STP, we calculated their magnitude and phase by complex demodulation of the logarithmic signals $10 \log_{10} P_v$ and $10 \log_{10} P_u$ (defined in Eq. 6). We took pitch-scaled frames of the signal, as for the PSHF ($N = 4T_0$, Hanning window w), and extracted the first harmonic, f_0:

$$\dot{P}_y(p) = \frac{10 \sum_{n=0}^{N-1} w(n) \exp\left(\frac{-j8\pi n}{N}\right) \log_{10} P_y(p + n - \frac{N}{2})}{\sum_{n=0}^{N-1} w(n)} \tag{7}$$

which provided the outputs $\dot{P}_v(p)$ and $\dot{P}_u(p)$ as complex Fourier coefficients, rather than as reconstructed single-harmonic signals. Implicit in the demodulation analysis is the assumption that the turbulence-noise source is multiplied by some signal that is related to the vibration of the vocal folds. Thus, by rejecting the higher harmonics, we can take this model as a first order approximation, and extract reliably the phase of the principal mode, that at the fundamental frequency.

The modulation amplitudes are shown in Figure 8 (top) and the relative phase (bottom). The modulation phases, which continually rotate at approximately the fundamental frequency f_0, are unwrapped and then subtracted from each other to form the phase difference between the modulation of the harmonic component and the modulation of the anharmonic component, as plotted (bottom). The degree of modulation of the harmonic part (Fig. 8 top, thick line) varies considerably during the vowel and the transition, but is more consistent during steady frication. The modulation amplitude is proportionately similar in the vowel and the fricative, and reaches its maximum value right at the transition into the fricative ($\sim$ 400 ms). It has minima at the points of weak voicing (around 520 ms and 640 ms), but otherwise grows in the fricative towards a steady value of approximately 6 dB. In contrast, the modulation of the anharmonic component is relatively constant throughout, although it is slightly higher at about 3 dB in the steady fricative. There are no clear trends in the vowel; in the fricative, it is arguable whether or not the dips following the points of weak voicing (550 ms and 690 ms) are significant, although quieter phonation might be expected to cause a reduction in the subsequent modulation.

The phase difference (see Figure 8 bottom), however, gives a more clear-cut picture. During the vowel, the phase difference between the two sets of modulation coefficients is approximately zero, but it changes abruptly at the transition towards a markedly different equilibrium c. −130°. We can calulate the mean phase more precisely by considering a series of unit vectors, each with its argument set equal to the instantaneous phase difference, θ:

$$\theta(n) = \arg\left(\frac{\dot{P}_u(n)}{\left|\dot{P}_u(n)\right|} \frac{\dot{P}_v^*(n)}{\left|\dot{P}_v(n)\right|} \right), \tag{8}$$

where $\dot{P}_v^*$ is the complex conjugate of $\dot{P}_v$, and $\dot{P}_y / \left|\dot{P}_y\right| = \exp\left(j \arg\left(\dot{P}_y\right)\right)$ is the unit vector with the same phase as the modulation coefficient $\dot{P}_y$, for any y. To avoid phase wrapping errors, unit vectors were used to average the phase in a mathematically-consistent circular algebra. Thus, the (unweighted)

time-averaged phase, with its standard deviation, is:

$$\langle\theta\rangle = \arg\left(\vec{e}_\theta\right) \pm \sqrt{\frac{\sum_{n=1}^{S} \left|\exp\left(j\theta(n)\right) - \vec{e}_\theta\right|^2}{S-1}}\,, \qquad (9)$$

in radians, where S is the number of sample points, and the mean unit vector $\vec{e}_\theta$ is:

$$\vec{e}_\theta = \frac{\sum_{n=1}^{S} \exp\left(j\theta(n)\right)}{S}\,. \qquad (10)$$

For token #2 in Figure 8 (bottom), $\langle\theta\rangle = -2° \pm 20°$ during the vowel (40–370 ms), and $-128° \pm 8°$ during the fricative (700–1000 ms) during the fricative. This marked difference suggests that more than one voiceless source is in action. The finding is not news in itself yet, as a positive result, it can be used to explore variations in the source interaction quantitatively.

4.4 Using EGG as a reference signal

In order to tell which component is causing the change in the phase difference, we sought to relate the phases to some independent measurement of the glottis. An ideal reference signal would be the glottal waveform itself, but for practical purposes, the glottal area or its electrical impedance, which can be obtained using an EGG, may be used. Using the coefficient of the EGG signal at f_0, $\dot{L}_x(n)$, we compute the phases of the components:

$$\phi_v(n) = \arg\left(\frac{\dot{P}_v(n)}{\left|\dot{P}_v(n)\right|}\frac{\dot{L}_x{}^*(n)}{\left|\dot{L}_x(n)\right|}\right)\,, \qquad (11)$$

$$\phi_u(n) = \arg\left(\frac{\dot{P}_u(n)}{\left|\dot{P}_u(n)\right|}\frac{\dot{L}_x{}^*(n)}{\left|\dot{L}_x(n)\right|}\right)\,. \qquad (12)$$

Ignoring the effect of phase wrapping, the phases can be subtracted to give Eq. 8: $\theta = \phi_u - \phi_v$.

Using the above method on the synthetic signals from Section 2.2, we estimated the phase offset β for each of its eight specified values (0°, 45°, 90°, etc.) at three HNRs (20, 10 and 5 dB). All modulation phases measured from the decomposed synthetic signals were within 5° of their specified values. The mean error was less than 1° and the inter-measurement standard deviation was 2°. There were no noticeable differences across the different HNR levels, except perhaps a slight trend in the (much higher) intra-measurement deviations, which were 15°, 13° and 13°, respectively.

Figure 9 contains the phase trajectories of the two components for another [ɑz:] token, #3, spoken by subject PJ, which do not exhibit the overshoot phenomenon that we saw earlier (Fig. 7 top). Both phases hover close to +90° initially. The harmonic component is perturbed near the transition, returning to approximately the same value for the fricative, except when it strays as voicing momentarily falters (between 1300 ms and 1430 ms).

The anharmonic component shows greater variability, but approaches an equilibrium value after the transition that is distinctly offset from the average during the vowel. The change noted in $\langle\theta\rangle$ thus appears to be due primarily to changes in ϕ_u, signalling a change in source mechanism for the unvoiced component. We expect that the anharmonic component during the vowel is due to a slight breathiness, i.e. turbulence noise generated in the vicinity of the glottis, and that during the following [z:], the anharmonic component is primarily due to turbulence noise generated downstream of the tongue-tip constriction. The step change in ϕ_u at the vowel-fricative transition therefore corresponds to a change in source location. This effect would predict that the amount of phase change should depend on the fricative's place, which we will investigate in Section 5. It should be noted that a phase difference of approximately zero could as easily be the product of perturbation errors (e.g. from jitter and shimmer) in the processing as of an in-phase modulated noise source. Nevertheless, examination of the time-series signals for the harmonic and anharmonic components for over twenty examples gives us confidence that the STP, as a summary of signal amplitude (or envelope), contains useful information about the sources.

5 Results

5.1 Sustained fricatives

The magnitude and phase of the modulation coefficients were determined for 10 fricative tokens that included seven different places of articulation. All of the tokens were similarly pitched at $f_0 = 120 \pm 5$ Hz, and sustained by subject PJ for at least 4 s, of which a steady section of approximately 1 s duration was analyzed. For some cases, the section analyzed included a part of the contextualizing vowel; for others, only the fricative was included. The PSHF was used to decompose each example, and modulation coefficients of the harmonic and anharmonic components were calculated, as described in Section 4. Finally, the coefficients were averaged over the fricative, excluding periods of devoicing, vowel-fricative transitions and two

pitch periods from either end of the section. The time-averaged magnitudes and phases are plotted in Figure 10. The points plotted on the vertical grid lines were all from steady regions of voicing, whereas those adjacent suffered an interruption in voicing.

As mentioned in Section 4.3, the magnitudes (Fig. 10, top) were all halved by the low-pass effect on signal power of the windowing, which was adjusted accordingly for each measurement to allow comparisons between harmonic and anharmonic STP, and across different phonemes. The magnitude of the modulation of the harmonic components (thick) is 3 ± 1 dB and, in all but one case, is greater than that of the anharmonic components (thin). The anharmonic modulation magnitudes were equally variable, but ranged from almost zero in the bilabial fricative [β] to 2 dB in [z] (the same as that of the harmonic modulation).

The phase of the modulation coefficients was referred to the EGG signal by subtracting the phase of its f_0 component, as before. Care had to be taken in aligning pitch, power (STP) and phase vectors in the analysis, but the difference between using the pitch extracted from the acoustic signal versus that from the EGG was found to be negligible. The unweighted-mean values are plotted in Figure 10 (bottom) with error bars indicating one standard deviation (±1 s.d.), time-averaged over the appropriate portion of the token. Of the two components, the harmonic's results showed greater consistency within each phase measurement; across measurements, these values were all in the vicinity of $+100° \pm 20°$. The anharmonic phases, although more variable, were all distinct from their harmonic phases, except for [ʕ]. Moreover, where the transition from the vowel was included in the analysis segment, a clear step was seen in the time series of the anharmonic modulation phase.

The phase of the modulation of [β]'s anharmonic component had the largest variance, which was related to the unusually small amount of modulation and rendered it most susceptible to interference from disturbances. Since the anharmonic modulation in [β] was therefore poorly correlated with the EGG, we shall ignore this phoneme in subsequent evaluation. For the remaining anharmonic phase data, there were two notable trends: (i) the mean phase increased as the place of constriction moved in a posterior direction, and (ii) so did the variance. The systematic change of phase with place seems worth further investigation, although we might well expect the phase to depend also on f_0. Any delay in the speech production system, such as the propagation time from the lips to the microphone, would add a phase term that increased linearly with f_0, its gradient dependent on the amount of delay. In the following section we investigate the relationship between the pitch and anharmonic phase during sustained fricatives that contain changes in f_0, and attempt to identify the cause of any delays.

5.2 Pitch glides

When using spot measurements of phase for determining delay times, the main concern is that phase wrapping may occur, e.g. a phase reading of 420° might be misinterpreted as only 60°, or vice-versa. The number of cycles is important because long delays, i.e. greater than a period, inherently entail phase-wrapping. A simple test for phase wrapping can be carried out by altering the fundamental frequency f_0 and by noting the phase changes. A few spot measurements can be made or, more dependably, a continuous measurement during a pitch glide. For a constant delay τ_u, the phase is simply a linear function of frequency:

$$\phi_u = 2\pi\tau_u f_0 + \beta\,, \tag{13}$$

where β is the phase offset between the actual modulating signal, whatever it may be, and the EGG signal. The phases ϕ_u and β can take any real value, although in our initial measurements they lie in the range ±180°. Hence, provided other independent variables remain unaltered, the gradient of the phase with respect to frequency provides an absolute estimate of τ_u, the delay duration for a given phoneme.

Subject PJ was asked to sustain a fricative during a smooth pitch glide sandwiched between two notes about a perfect fifth apart. That is, a constant-f_0 fricative was held for at least 1 s, then f_0 was increased steadily to approximately $1.5f_0$ over a similar period, and finally the fricative was held at the higher note of about $1.5f_0$ for at least another second, taking about 5 s in total. Recordings were also made of descending pitch glides.

For all of the tokens analyzed, the time series of the anharmonic modulation phase showed a definite correlation with the extracted f_0, and both parameters exhibited distinct equilibria at the end conditions, which were connected by a gradual transition. The relationship between f_0 and the phase ϕ_u can be seen more clearly by plotting them against each other, independently of time. Thus, Figure 11 is a scatter diagram of the anharmonic STP modulation phase versus fundamental frequency for the sustained fricative [z:], during a descending pitch glide.[5] In this example, the

points lie roughly along a diagonal line, in the range $\pm 45°$, except for a few stray excursions that occured at transitions or near a singularity, where the modulation amplitude was almost zero. There is a higher density of points at either end of the trajectory line due to the period of constant pitch before and after the frequency ramp. The deviation from this line, $\sigma \approx 10°$, is of the same order as the deviation of the (constant-f_0) sustained fricatives considered earlier. Owing to the integer quantization of the extracted pitch period (in sample points), the fundamental frequency values also exhibit quantization, which explains why the data points lie on a set of vertical lines.

The best-fit line (thick solid line in Fig. 11) was calculated for the plotted data points by a least-mean-squares regression and provides good general agreement. The line's gradient provides an estimated delay time of $\tau_u \approx 3.8\,\mathrm{ms}$, and the intercept with the y-axis at $f_0 = 0$, was $\beta \approx -170°$. Regression lines were also calculated for two other examples: [z:] ascending and [ʒ:] descending. The lines for [z:] are within 10° of each other for the ranges of f_0 measured, although their gradients differ, which suggests that some other factor may have influenced these results. The line for a descending [ʒ:] is set apart from those for [z:], but has a similar gradient, particularly to that of the descending [z:].

The values of β and τ_u for all three cases are listed in Table 2, with the mean values of the f_0-glide endpoints. The difference between the two descending fricatives [z:] and [ʒ:] was as expected in both direction and scale, yet there was a considerable discrepancy between the values calculated for the ascending and descending [z:], which was exacerbated by the extrapolation to $f_0 = 0$. Given that the propagation time for an acoustic wave from the lips to the microphone is 2.9 ms ($r = 1\,\mathrm{m}$, $c_0 = 343\,\mathrm{m/s}$, room temperature, dry air) and acoustic propagation in the tract would take about 0.5 ms ($l = 16.5\,\mathrm{cm}$, $c_0 = 359\,\mathrm{m/s}$, body temperature, saturated air), the times derived from the gradient are of an appropriate order of magnitude. The zero-frequency phase offset β, despite these errors, corresponds to a point between one-half and three-fourths of the way through the open portion of the glottal cycle. We shall speculate about potential interpretations of the coincidence of this timing relationship with the maximum glottal flow in the following section. For fricatives showing a higher variance, the scatter plots are less informative. Critically, no phase wrapping of the modal trajectories took place for any of the fricatives examined, which validates the order of our earlier phase measurements.

6 Discussion

6.1 From phase to delay

We would like to be able to convert the reported phase values into delay times in order to relate a peak in the acoustic response to the event that caused it. The glottal closure is commonly assumed to give the principal acoustic excitation of the vocal tract. The harmonic component $v(n)$ should then consist primarily of the vocal tract response to that excitation. The smoothed STP of $v(n)$ has a peak every cycle that is slightly delayed with respect to the instant of excitation, and further delayed due to the acoustic propagation time from the glottis to the microphone in the far field. We computed its phase ϕ_v with respect to the peak of the fundamental component of the EGG signal. To refer it instead to the moment of closure of the vocal folds, we subtract $\alpha = \arg\left(\dot{L}_x\right)_{\mathrm{cl}}$; to convert this phase to a time delay, we divide by the instantaneous fundamental frequency:

$$\tau_v = \frac{\phi_v - \alpha}{2\pi f_0}\,, \tag{14}$$

where ϕ_v is defined by Eq. 11. The anharmonic component $u(n)$ consists primarily of the vocal tract response to the noise excitation. We wish to convert ϕ_u to a time delay also, but it is not clear whether we should refer ϕ_u to the same instant of closure of the EGG signal. If we use the same angle α as in Eq. 14, we are effectively assuming a model of the modulation mechanism, namely that the peak amplitude of the turbulence noise source is evoked by the excitation originating from the instant of glottal closure. We wish instead to deduce the mechanism controlling the modulation, by using the phase difference expressed as a time delay. Therefore, to refer the phase to an unknown point in the EGG signal, we subtract the angle β:

$$\tau_u = \frac{\phi_u - \beta}{2\pi f_0}\,. \tag{15}$$

where ϕ_u is defined by Eq. 11. For our initial discussions, we set $\beta = \alpha$.

Figure 12 shows a set of four synchronous time-series signals during the fricative [z:] sustained by subject PJ, which are (from top) recorded EGG $L_x(n)$, recorded sound pressure $s(n)$, and the decomposition into the harmonic and anharmonic signals, $v(n)$ and $u(n)$. The dashed lines around the harmonic and anharmonic components represent their envelopes (i.e.,

$\pm 2\sqrt{P_v}$ and $\pm 2\sqrt{P_u}$). The EGG measures the time-varying (high-pass filtered) part of the trans-glottal conductance, which is at a maximum when the glottis is closed. It shows a sharp rise at the instant of closure, occuring at around -0.4π ($-72°$), with respect to the EGG signal's fundamental component, whose phase is indicated by the upper abscissa in Fig. 12. This phase offset is slightly less than a quarter of a cycle, because of the long open portion and the abruptness of the closure. Although the phase may change slightly throughout the recorded corpus and for subjects other than PJ, the value of $\alpha = -0.4\pi$ shown here is used in all cases to refer the harmonic component to the same instant of the EGG signal.

Through a separate study (Shadle et al. 1999), we obtained magnetic resonance imaging (MRI) data for subject PJ, saying [$\mathrm{p^h}$ɑsi]. Combining these with articulatory phonetics, we were able to estimate the constriction location for each phoneme. Distances along the vocal tract were measured from the glottis, and the position of the teeth was estimated in relation to the lips and the hard palate (upper) or tongue body (lower). Table 3 lists all the constriction-teeth distances, which agree closely with Table I in Narayanan et al. (1995). For the breathy vowel [$\mathrm{ɑ^h}$], the place of greatest constriction was assumed to be the glottis.

Ideally, we would like to characterize each phoneme by two distances: from glottis to place of constriction, and from constriction place to the location of turbulence noise generation. Different aspects of sound generation take place over these two 'paths'. However, while for some fricatives it is well known that noise generation is highly localized at the teeth (e.g., [s, ʃ, z, ʒ]), for others the noise source appears to be distributed, for instance, along the hard palate for [ç] (Shadle 1991). The distance from the constriction to the source location is thus less precisely known for some fricatives. All delays are therefore calculated using the constriction-teeth distances given in Table 3. These values were used for all three subjects, regardless of minor inter-subject variation in physical dimensions. Although women's vocal tracts are generally shorter than those of men, most of the difference is in the pharynx. Since for LJ and SB we are dealing with distances from within the oral cavity to the teeth, the variation is considered negligible. Although this part of the procedure is crude compared with the signal processing, it enables us to visualize our results in a way that has greater physical meaning. Bearing in mind that the teeth will not necessarily be the source location in all cases, we can nevertheless interpret trends and make order of magnitude calculations to help indicate the aero-acoustic processes that are likely to be operating.

The delays calculated for the voiced fricatives of three subjects are plotted against place of articulation in Figure 13, including one breathy [ɑ]-vowel (PJ). For reference, the lip-microphone propagation time is shown as a dashed horizontal line, $\tau_R = 2.9\,\mathrm{ms}$ for a microphone at 1 m (speed of sound $c_0 = 343\,\mathrm{m/s}$). In Figure 13 (top), the delay times τ_v are all greater than the acoustic propagation delay, as expected. The additional delay, the reverberation lag, is reasonably consistent across phonemes, showing a mean value of 1.3 ms and no significant trend. In contrast, τ_u (Fig. 13, bottom) is generally below τ_R. Since the largest portion of these delays is, in fact, the wave propagation time from the lips to the microphone (which is obviously identical for both components), any variations in the delay are attributable to other causes. Such causes include jitter/shimmer effects, changes in glottal waveform, changes in vocal-tract configuration, the measurement noise on the data, processing errors, and actual changes in the source characteristics. However, before we attempt to interpret the anharmonic τ_u results, let us consider the physical mechanisms that could lead to modulation of the frication source, as has been observed.

6.2 Travel times

For all voiced fricatives, the path that the flow perturbation must take from glottis to far-field microphone can be divided into three sections: from glottis to constriction exit; from constriction exit to the principal location of turbulence noise generation; thence to the microphone. The first two paths are the most important with regard to the mechanism of noise modulation.

During phonation, the pulsing jet of air exiting from the glottis generates sound and sets up vortical motion. The sound wave travels downstream at the speed of sound; the vortices convect at the order of the mean flow velocity, which is much slower than the speed of sound c_0 (Barney et al. 1999). The effects of phonation therefore traverse the first section of the path in two different ways, with two different travel times. The longer that section is, i.e. the more anterior the constriction, the bigger the discrepancy in time will be.

The travel time for a sound wave over this first glottis-to-constriction path of length l_1 can be estimated as $\tau_1|_{ac} = l_1/c_0$. Values are shown in Table 4

computed for three different l_1 values ($c_0 = 359\,\mathrm{m/s}$). The convective travel time is estimated as $\tau_1|_{co} = l_1/(V/2)$. A minimum and maximum convective velocity are computed using volume velocities of 200 and $600\,\mathrm{cm}^3/\mathrm{s}$, and an average cross-sectional area through the back cavity of $5\,\mathrm{cm}^2$. It is clear from the values shown in the Table that even the lower of the convective delay estimates (co_2) is two orders of magnitude higher than the measured delays. Such delays would be easily observable at any transition, and would in particular lead to extensive phase wrapping on the pitch glides. Further, we observe longer delays (longer by approximately 1 ms) for a more posterior place, whereas a convective mechanism for path 1 would mean that delays would shorten by 50 to 150 ms. Therefore we conclude that the aspect of phonation that modulates the noise travels at the speed of sound over path 1.

The second path extends from the constriction to the principal location of turbulence noise generation. The flow velocity increases in the constriction; at the exit, a turbulent jet forms. The self-noise (from mixing) of the jet is relatively weak for vocal-tract dimensions and flow rates but, whatever obstacle the jet encounters (whether the palate or the teeth), additional turbulence noise is generated that is louder (and can be much more localized). If the jet emerging from the constriction is pulsing, the turbulence noise generated by it will likewise fluctuate, but an acoustic field can also influence the formation of turbulence (Crow and Champagne 1971). We could further consider whether an acoustic field could influence not only the jet structure, but the sound generation where it impinges on the obstacle.

For path 2, we can again make order-of-magnitude estimates of the travel time at acoustic and convective velocities. We estimate l_2 to be the constriction-teeth distance, although we expect that the teeth do not act as the obstacle in all these cases. Again, two values of l_2 are chosen that correspond to the two values of l_1, that is, result in the same vocal tract length in both cases. The acoustic delay is then computed as $\tau_2|_{ac} = l_2/c_0$, as shown in the table. For the convective delay, V is recomputed using a typical constriction area of $0.1\,\mathrm{cm}^2$ rather than the $5\,\mathrm{cm}^2$ used earlier. The same minimum and maximum volume velocites are used, giving much higher values of V.

From Figure 13 (bottom), lengthening l_2 from 2 to 5 cm actually increases the delay by approximately 0.7 ms. This is consistent with the convective delay computed using the maximum convective velocity (column co_2 in Table 4). If travel times were at speed of sound in both paths, there would be virtually no difference in the delay with place. Therefore, the second path must involve some mechanism that convects.

6.3 Source modulation mechanisms

What theoretical models exist that describe the modulation mechanism itself? Most of the methods in the literature, summarized in the Introduction, incorporate modulation by a parameter related to glottal flow, such as the instantaneous component of the volume velocity at the constriction exit, but do not allow for a non-acoustic mechanism, i.e. for propagation velocities other than the speed of sound. The differences with place that we observe in the phase of the anharmonic component are not consistent with models depending only on acoustic propagation.

We have not so far discussed the extensive literature examining interaction of the glottal waveform with the vocal-tract driving-point impedance. Rothenberg (1981) showed, theoretically and by inverse-filtering speech, that the first formant frequency F_1 affects the degree of skewing of the glottal waveform U_G: the vowel [ɑ], with its high F_1, has a more skewed U_G (peak U_G occuring later in the glottal cycle) than does [i], with low F_1. Since all of the English voiced fricatives have lower F_1 than [ɑ], the peak U_G is predicted to shift earlier in the cycle during [ɑF], which was borne out by Bickley and Stevens' results (1986) for consonantal constrictions at the lips. Nevertheless, though such a mechanism could perhaps explain why the phase difference changes during the vowel-fricative transition, it does not explain the amount of change we observe (ranging from 40° to 150°) nor the difference with place, which should affect F_2 and higher formants rather than F_1.

Crow and Champagne (1971) showed that acoustic excitation applied to air in a duct upstream of the jet nozzle could induce an orderly structure in the jet wake, with a preference for $St = fD/V = 0.30$. Such a structure appears when the acoustic velocity is greater than 1 % of the mean flow speed V at the nozzle exit (nozzle diameter D). The turbulence noise spectra show that the forcing has the effect of suppressing background noise and enhancing noise at frequencies near the forcing fundamental and its harmonics.

We cannot compare all aspects of Crow and Champagne's results to ours because the relevant vocal-tract parameters cannot be measured accurately enough. However, we estimate that Strouhal numbers for voiced fricatives range from 0.3 to 0.9, based on $f = f_0$, a typ-

ical constriction diameter D, and the volume velocities U used in Table 4. The forcing takes some (unspecified) time to alter the shape of the jet; any change in the jet travels downstream at its convection velocity. We conjecture that the sound generation mechanism with which we are chiefly concerned, that of the jet impinging on an obstacle, would, in the presence of the 'forcing function' of phonation at f_0, exhibit nonlinear emphasis of f_0 and its harmonics, similar to the free jet spectra shown by Crow and Champagne. Any change in f_0 would affect the noise generated after a delay, related to the convection velocity and the distance from constriction to obstacle. Their results provide a plausible mechanism for the modulation of voiced fricatives, but do not help us to estimate β, the angle that determines the phase of the glottal cycle to which we should refer the modulation of the anharmonic component. Nevertheless, we can place some bounds on β's range of variation.

6.4 Interpretation

Up to this point, we have set $\beta = \alpha = -72°$. However, this produced delays shorter than the acoustic propagation time from lips to microphone, i.e. $\tau_u < \tau_R$. This is not possible since if any part of the path is traveled at convection velocity, the delay will be increased. Therefore $\beta < \alpha$, i.e. β is more negative than α. Yet β has a lower bound, since otherwise we would observe phase wrapping during the pitch glides. (For the interval of a perfect fifth used here, the lower bound is -6π.) We thus have strong bounds on β: $-(3 \times 360)° < \beta < -72°$. In addition, we can compute the angle that would make the minimum τ_u just equal to the acoustic propagation of 2.9 ms: $\beta \lesssim -175°$.

The pitch glide data produced estimates of β that ranged from -120 to $-180°$, as presented in Table 2. The estimates so derived must be treated with caution for two reasons: they are based on one subject and only three glides, and the fitted lines are used to extrapolate an intercept value. Thus any variation in the glide itself will be magnified in the intercept estimate. By modifying the best fit lines to the pitch glide results, using one standard deviation to give the worst case gradients, we get a range of $-200° < \beta < -100°$. These weak bounds for the range of β, together with the stronger bounds given above, predict that β in Eq. 15 should lie within the range: $-200° \lesssim \beta \lesssim -175°$. Taking $\beta = -175°$ would effectively add 2.4 ms to the delays shown in the lower half of Figure 13.

While it is clear that modulation of the anharmonic component varies with place, we can do no more than speculate that the acoustic-convective theory of sound production for the fricative component in voiced fricatives is the most likely, whose mechanism can be described as follows. A pulsed flow is emitted from the glottis into the vocal tract. Sound waves propagate down the vocal tract towards the constriction; at the constriction, the flow forms a jet, developing turbulence as it travels downstream. The temporal and spatial characteristics of the mixing flow are strongly influenced by the intersecting sound waves, inducing synchronous pulses of turbulence; the pulsed turbulence and entrained vortices convect downstream. When the jet encounters an obstacle (such as the teeth), a new source is generated that is pulsed at f_0 and efficiently radiates sound. The sound source at the obstacle excites the vocal tract; sound radiated from the lips propagates into the far field.

Assuming this to be the case, the increasing variance in Section 5.1 might be explained by three possible causes. First, the exact shape and location of the constriction may vary more for more posterior places, as the articulators become larger and are less finely controlled (e.g., tongue dorsum relative to tongue apex). Second, variations in convection velocity would make a larger contribution for the more posterior fricatives where the vorticity has further to travel before reaching the obstacle. Third, the obstacle upon which the turbulence impinges is likely to extend further in the direction of flow, producing a more distributed source for constrictions nearer to the glottis.

7 Conclusion

In this paper, we have used the pitch-scaled harmonic filter (PSHF) on voiced fricatives to decompose them into harmonic and anharmonic components. The amplitude of the components was represented by their short-time power, which exhibited modulation at the fundamental frequency f_0. The relative phase of the modulation of the two components changes rapidly at a vowel-fricative transition, settling near an equilibrium that depends on the fricative's place of articulation. The subjects were recorded uttering fricatives at a range of places. The findings of this article support the suggestion that the aero-acoustic mechanism of fricative sound production is modified by voicing, due to the powerful effect of upstream acoustic disturbances as they intersect the jet (Crow and Champagne 1971).

Tests of our PSHF algorithm on synthetic signals

confirmed that modulation was not a signal processing artifact, and predicted improvements to the SER greater than 5 dB on the harmonic part and of the HNR plus 5 dB on the anharmonic part. The algorithm was then applied to give a plausible decomposition of the recorded utterance [p$^{\text{h}}$ɑzɑ], successfully separating simultaneous parts of voiced and unvoiced speech. Inspecting the reconstructed time series, we observed the time-varying interaction of sources in the voiced fricative [z:], manifested as pulsing of the unvoiced component. Using the STP to approximate the signal envelopes, we derived an objective and quantitative method for measuring the magnitude and phase of the pulsation by complex demodulation. The phase difference between the modulation of the harmonic and anharmonic parts revealed two distinct states in the vowel-fricative transition [ɑz:]. Referring the phase values to the EGG provided better fidelity in the modulation analysis and allowed us to attribute the change in state to the anharmonic component, which corresponded to a change in the unvoiced source location. The phase change decreased as the place of the constriction moved posteriorly, which was verified on a second subject (LJ).

A set of f_0 glide experiments showed that the phase, as a function of f_0, behaves almost entirely like a constant place-dependent delay. It is tempting to speculate further about the role of the observed phase differences in the categorical perception of voiced fricatives, particularly in opposition to aspiration noise, but we have found scant empirical evidence in the literature to support these claims. In perceptual tests on synthetic signals, Hermes (1991) found that the perception of noise bursts is affected by their phase relative to voicing; out-of-phase noise is distinguished from the voicing component, whereas synchronous bursts are assimilated.

In summary, we have used a pitch-scaled harmonic filter to decompose voiced fricatives into harmonic and anharmonic components. The different phase of the envelopes of these components led us to vary place and f_0 systematically in order to determine the mechanism controlling the modulation. We have shown that a plausible explanation is that the acoustic signal generated at the glottis induces a structure in the jet emerging from the constriction, and thus alters the noise generated by the jet as it impinges on an obstacle. Further practical experiments using dynamic physical models should be conducted to establish whether this explanation is correct. The second non-acoustic path that accounts for the variation of phase with place has not been incorporated into speech synthesis models until recently (Sinder 1999). It would be instructive to ascertain whether Sinder's model predicts the phase changes we observed. It would also be useful to explore inter-subject variations and the robustness of phase changes to changes in f_0, effort and speaking style. Finally, the phase difference between harmonic and anharmonic components, which changes suddenly in the vowel-fricative transition, may well be perceptually important and should be investigated.[6]

Acknowledgments

This paper is based on a talk presented at the 2nd International Conference on Voice Physiology and Biomechanics, Berlin, Germany, 12–14 March 1999. The authors would like to thank Phil Nelson and Anna Barney, both at ISVR, University of Southampton, and Celia Scully, formerly at University of Leeds, for helpful discussions. The authors would also like to thank Dirk Michaelis, Drittes Physikalisches Institut, Göttingen, and Dan Sinder, Lincoln Laboratory, MIT, for their helpful comments on an earlier version of this manuscript.

[1] There is no reason why, in theory, a number of periods other than four may be not used, but we have not tested any alternatives. However, we believe that the current value, which has a time-frame comparable to others (e.g., Frazier et al. 1976), offers a reasonable compromise between adaptability and ideal PSHF performance for speech signals.

[2] In a similar study, the PSHF performance was evaluated with three kinds of perturbation: jitter, shimmer and constant-variance additive noise. Although those tests were at a different pitch ($f_0 = 130.8$ Hz), the performance at matching conditions was unaffected.

[3] Incidentally, repeating the process with the prescribed pitch values showed that our using the noisy values had little effect on the anharmonic performance, which was degraded by 0.4 dB in the worst case. The observed decline in the harmonic performance with increasing noise, though, was entirely due to the effect of noise on the estimated pitch, which would otherwise have kept η_v pinned at 5.4 dB and 5.6 dB for all constant and modulated noise tests, respectively.

[4] Note that the STP can also be computed in a pitch-scaled way, but there is little advantage from this minor adjustment to the roll-off frequency, for the range of f_0 values within each token.

[5] Every one in ten points has been plotted, so the values have been effectively sampled at 4.8 kHz.

[6] Further information can be found on the internet, including Matlab script (`.m`) files of the algorithm, a data (`.dat`) file containing the LPC coefficients used in Section 2.2 and sound (`.wav`) files of examples used in this paper: http://www.isis.ecs.soton.ac.uk/research/projects/nephthys/.

References

Barney, A. M., C. H. Shadle, and P. O. A. L. Davies (1999). Fluid flow in a dynamic mechanical model of the vocal folds and tract. I. Measurements and theory. *J. Acoust. Soc. Am. 105*(1), 444–455.

Bickley, C. and K. N. Stevens (1986). Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies. *J. Phon.* (14), 373–382.

Bretthorst, G. L. (1988). *Bayesian Spectrum Analysis and Parameter Estimation*. Lecture Notes in Statistics. Berlin, FRG: Springer-Verlag.

Coker, C. H., M. H. Krane, B. Y. Reis, and R. A. Kubli (1996). Search for unexplored effects in speech production. *Proc. ICSLP '96, Philadelphia, PA 14*(6), 415–422.

Crow, S. C. and F. H. Champagne (1971). Orderly structure in jet turbulence. *J. Fluid Mech. 48*, 547–591.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton.

Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception* (2nd ed.). Berlin: Springer-Verlag.

Flanagan, J. L. and L. Cherry (1969). Excitation of vocal-tract synthesizers. *J. Acoust. Soc. Am. 45*(3), 764–769.

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

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

Hermes, D. J. (1991). Synthesis of breathy vowels: some research methods. *Speech Comm. 10*(5-6), 497–502.

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

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am. 67*(3), 971–995.

Klatt, D. H. and L. C. Klatt (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am. 87*(2), 820–857.

Laroche, J., Y. Stylianou, and E. Moulines (1993). HNS: Speech modification based on a harmonic + noise model. *Proc. IEEE-ICASSP 93*(2), 550–553.

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

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

Narayanan, S., A. Alwan, and K. Haker (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *J. Acoust. Soc. Am. 98*(3), 1325–1347.

Pelorson, X., G. C. J. Hofmans, M. Ranucci, and R. C. M. Bosch (1997). On the fluid mechanics of bilabial plosives. *Speech Comm. 22*, 155–172.

Qi, Y. and R. E. Hillman (1997). Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *J. Acoust. Soc. Am. 102*(1), 537–543.

Rife, D. C. and R. R. Boorstyn (1974). Single-tone parameter estimation from discrete-time observations. *IEEE Trans. Inf. Theory 20*(5), 591–598.

Rothenberg, M. R. (1981). Acoustic interaction between the glottal source and the vocal tract. *Vocal Fold Physiology,* eds. K. N. Stevens and M. Hirano, Univ. of Tokyo Press, 305–328.

Scully, C. (1990). Articulatory synthesis. In W. J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 151–186. Kluwer Academic.

Scully, C., E. Castelli, E. Brearley, and M. Shirt (1992). Analysis and simulation of a speaker's aerodynamic and acoustic patterns for fricatives. *J. Phon. 20*, 39–51.

Serra, X. and J. Smith (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition. *Comp. Mus. J. 14*(4), 12–24.

Shadle, C. H. (1991). The effect of geometry on source mechanisms of fricative consonants. *J. Phon. 19*(3-4), 409–424.

Shadle, C. H., M. A. S. Mohammad, J. N. Carter, and P. J. B. Jackson (1999). Multi-planar dynamic Magnetic Resonance Imaging: New tools for speech research. *Proc. Int. Cong. on Phon. Sci., San Francisco, CA 1*, 623–626.

Silva, F. M. and L. B. Almeida (1990). Speech separation by means of stationary least-squares harmonic estimation. *Proc. IEEE-ICASSP 2*, 809–812.

Sinder, D. J. (1999). *Speech synthesis using an aeroacoustic fricative model*. Ph. D. thesis, Rutgers Univ., New Brunswick, NJ.

Sondhi, M. M. and J. Schroeter (1987). A hybrid time-frequency domain articulatory speech synthesiser. *IEEE Trans. ASSP 35*(7), 955–967.

Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *J. Acoust. Soc. Am. 50*(4, Part 2), 1180–1192.

Yegnanarayana, B., C. R. d'Alessandro, and V. Darsinos (1998). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. SAP 6*(1), 1–11.

Table 1: PSHF performance versus HNR for synthetic signals with constant and modulated noise ($\beta = \pi$); results are η_u (η_v) in dB.

HNR	Constant		Modulated	
∞	72.6	$(-\infty)$	72.6	$(-\infty)$
20 dB	25.2	(5.3)	25.4	(5.6)
10 dB	15.1	(5.2)	15.4	(5.5)
5 dB	10.1	(5.2)	10.2	(5.4)
0 dB	4.9	(5.1)	5.0	(5.2)
−5 dB	−0.0	(5.1)	−1.0	(4.2)

Table 2: The anharmonic delay τ_u, the offset phase β and the standard deviation σ about the corresponding regression line, for three f_0 glides by subject PJ.

Phoneme	f_0 (Hz)	τ_u (ms)	β (°)	σ (°)
[z:] ascending	125 → 175	2.8	−129	10
[z:] descending	111 ← 172	3.8	−169	11
[ʒ:] descending	121 ← 178	4.0	−154	22

Table 3: Estimated distance from the constriction to the teeth for sustained voiced fricatives by subject PJ, in cm.

Phoneme	v	ð	z	ʒ	ɣ	ʕ	ɑʰ
Distance	0.0	0.4	1.1	2.2	5.2	10.3	14.9

Figure 1: Flow diagram of the pitch-scaled harmonic filter (PSHF) algorithm. See text for explanation of the harmonic filter (HF), power interpolation (PI) and factor λ.

Table 4: Estimated travel times (ms) for /z/ ($l_1 = 14.6\,\text{cm}$, $l_2 = 1.1\,\text{cm}$), /ʒ/ ($l_1 = 13.5\,\text{cm}$, $l_2 = 2.2\,\text{cm}$) and /ɣ/ ($l_1 = 10.2\,\text{cm}$, $l_2 = 5.2\,\text{cm}$), by acoustic propagation *ac* or by convection *co*, using $U_1 = 200\,\text{cm}^3/\text{s}$ and $U_2 = 600\,\text{cm}^3/\text{s}$ for co_1 and co_2 respectively. The column under t_1 gives the travel times over path 1, and the first row under t_2 those for path 2. The nine values inside each sub-table are $t_1 + t_2$, rounded to two significant figures; those in bold face best match the measured data (see text).

/z/		t_2 (ms)			/ʒ/		t_2 (ms)			/ɣ/		t_2 (ms)		
		ac	co_1	co_2			*ac*	co_1	co_2			*ac*	co_1	co_2
t_1 (ms)		0.06	1.90	0.63	t_1 (ms)		0.08	3.0	1.0	t_1 (ms)		0.17	6.0	2.0
ac	0.38	0.44	**2.3**	**1.0**	*ac*	0.35	0.44	**3.4**	**1.4**	*ac*	0.27	0.44	**6.3**	**2.3**
co_1	690	690	690	690	co_1	640	640	640	640	co_1	490	490	490	490
co_2	230	230	230	230	co_2	210	210	220	210	co_2	160	160	170	160

Figure 3: Time series of the synthetic signal $s(n)$ with its constituent harmonic and anharmonic parts $v(n)$ and $u(n)$, the PSHF signal estimates $\hat{v}(n)$ and $\hat{u}(n)$, and the error $e(n)$, at HNR = 10 dB for (left) constant-variance noise, and (right) modulated noise with $\beta = \pi$. They are arranged, from top to bottom, thus: s, v, $\hat{v}$, u, $\hat{u}$ and e (anharmonic and error signals are double amplitude scale).

Figure 4: Time series, from #1 by subject PJ, of (top) the original signal $s(n)$, (middle) the harmonic component $\hat{v}(n)$ and (bottom, double amplitude scale) the anharmonic component $\hat{u}(n)$.

Figure 5: Spectrograms (5 ms, Hanning window, $\times 4$ zero-padded, fixed gray-scale) computed from the decomposition of #1 by subject PJ: (top) the original signal $s(n)$, (middle) the harmonic estimate $\hat{v}(n)$ and (bottom) the anharmonic estimate $\hat{u}(n)$.

Figure 6: A detailed view of the time series, from the vowel-fricative transition [-ɑz-] in #1 by subject PJ, of (top) the original signal $s(n)$, (middle) the harmonic component $\hat{v}(n)$ and (bottom, double amplitude scale) the anharmonic component $\hat{u}(n)$.

Figure 7: The short-time power (STP) calculated over the medium term (top, $M \approx 32\,\text{ms}$) and the short term (bottom, $M \approx 8\,\text{ms}$) for the decomposed components from $^{\#}2$ by subject PJ: (thick) harmonic P_v, and (thin) anharmonic P_u.

Figure 8: Modulation of the short-term STPs at f_0 using token $^{\#}2$ by subject PJ, plotted as magnitudes (top: harmonic, thick; anharmonic, thin) and the phase difference (bottom).

Figure 9: Phase of the harmonic (thick) and anharmonic (thin) modulation components for #3 by subject PJ, related to that of the simultaneously-recorded EGG signal.

Figure 10: Magnitude (top) and phase (bottom) of modulation coefficients, referred to the EGG signal, versus place of articulation for sustained fricatives [β, v, ð, z, ʒ, ɣ, ʕ] by subject PJ. Harmonic (•, thick line) and anharmonic (×, thin line) components were plotted with ($\pm 1\sigma$) error bars. Those measurements on vertical grid lines are for normal voicing; those adjacent (to the right), where a pair of measurements are shown, were taken from a section interrupted by devoicing.

Figure 11: Scatter plot of the anharmonic modulation phase versus fundamental frequency for the sustained fricative [z:] by subject PJ during a descending pitch glide, with its regression (thick solid line), and those of an ascending [z:] (thin solid line) and a descending [ʒ:] (thick dashed line).

Figure 12: Time series during a sustained [z:] by subject PJ: (from top) EGG signal L_x, sound pressure s, harmonic part v, and anharmonic part u.

Figure 13: Harmonic and anharmonic delay times, τ_v (top, Eq. 14) and τ_u (bottom, Eq. 15) respectively, versus distance of constriction from teeth, for subjects PJ ($\diamondsuit$), LJ ($\odot$) and SB ($\star$). The dashed line is the predicted lip-mic propagation delay τ_R, the thin solid line is the predicted total delay, and the thick solid line is the quadratic line of best fit.