

Amplitude modulation of turbulence noise by voicing in fricatives

Jonathan Pincas and Philip J. B. Jackson

Centre for Vision, Speech & Signal Processing,
University of Surrey, Guildford, GU2 7XH, UK
j.pincas@surrey.ac.uk

29th August 2006

Amplitude modulation of frication by voicing

The two principal sources of sound in speech, voicing and frication, occur simultaneously in voiced fricatives as well as at the vowel-fricative boundary in phonologically voiceless fricatives. Instead of simply overlapping, the two sources interact. This paper is an acoustic study of one such interaction effect: the amplitude modulation of the frication component when voicing is present. Corpora of sustained and fluent-speech English fricatives were recorded and analyzed using a signal-processing technique designed to extract estimates of modulation depth. Results reveal a pattern, consistent across speaking style, speakers and places of articulation, for modulation at f_0 to rise at low voicing strengths and subsequently saturate. Voicing strength needed to produce saturation varied 60–66 dB across subjects and experimental conditions. Modulation depths at saturation varied little across speakers but significantly for place of articulation (with [z] showing particularly strong modulation) clustering at approximately 0.4–0.5 (a 40–50% fluctuation above and below unmodulated amplitude); spectral analysis of modulating signals revealed weak but detectable modulation at the second and third harmonics (i.e., $2f_0$ and $3f_0$).

PACS numbers: 43.70.Bk, 43.72.Ar

I INTRODUCTION

English has voiceless and voiced fricatives at four places of articulation: postalveolar /ʃ,ʒ/, alveolar /s,z/, dental /θ,ð/ and labiodental /f,v/. These speech sounds are produced by forcing air through a narrow constriction in the oral cavity (Shadle, 1985), generating noise within the jet and, more importantly, at or along a physical obstacle downstream, such as the teeth (sibilants /ʃ,ʒ,s,z/) and lips (non-sibilants /θ,ð,f,v/), as shown in Figure 1.

Acoustic theory of noise generation from turbulence describes three types of source: monopole, dipole and quadrupole (Lighthill, 1952). Monopoles arise from velocity fluctuations injected into the soundfield, dipole sources result from turbulent flow impinging on a solid obstacle, and quadrupoles occur in regions of turbulence through self mixing. The intensity of these sources depends on the flow velocity as V^4 , V^6 and V^8 respectively (Lighthill, 1954).

In fricatives, flow at the constriction exit produces a monopole, then quadrupoles just downstream in the jet core and dipoles at the teeth or lips (Stevens, 1998). It is widely accepted that dipole sources dominate noise generation in fricatives (Stevens, 1971; Shadle, 1985, 1990), although some studies have considered a monopole component (Pastel, 1987; Stevens, 1998; Narayanan and Alwan, 2000).

Voiced frication has both glottal and fricative sources, which produces the familiar ‘buzzy’ quality. Yet it does not occur solely in voiced fricatives. Pincas (2004) recorded an average 16 ms of voicing overlapping frication at the vowel-fricative (/VF/) boundary of voiceless fricatives; Heid and Hawkins (1999) report 72% of voiceless fricatives having mixed excitation at the /VF/ boundary, according to an 8 ms minimum source overlap. Figure 2 shows a vowel transition into a *voiced* fricative. At transition, the formants move and fade; meanwhile, the high-frequency noise becomes prominent during the fricative segment.

The characteristics of voiced frication do not arise simply from the linear combination of independent sources. The articulatory, aerodynamic and acoustic conditions required by and resulting from the simultaneous production of glottal vibration and frication noise produce ‘mutual interaction effects’ (Pincas and Jackson, 2004): the presence of each source

causes the other to be changed. The focus of this paper, amplitude modulation (AM) of the frication component, is one such effect and can be seen as vertical striations in the spectrogram of Fig. 2(a). Other effects include mutual amplitude reduction (Stevens, 1971), changes in fundamental frequency of voicing (Lofqvist et al., 1989), and spectral changes both in the voicing component before, during and after frication (Lofqvist et al., 1995) and in the frication-noise component (Shadle, 1995).

A Amplitude modulation formulation

Modulation depth m is the aspect of AM studied here. It can be conceptualized as the fraction of the carrier signal by which the modulated signal varies, e.g., if $m = 0.5$, then the signal fluctuates by 50% above and below its original, unmodulated value. It is most often given in standard index form (in the range 0–1), but in the perceptual literature modulation depth is often quoted in dB: $20\log_{10}(m)$.

In AM, a carrier signal $w(n)$ is multiplied by a modulating signal $a(n)$ to produce the amplitude-modulated signal, $x(n) = w(n)a(n)$. With a periodic modulating signal, $a(n)$ takes the form of a d.c. term and fundamental sinusoid of frequency f_0 plus harmonics:

$$x(n) = w(n) \left[1 + \sum_{h=1}^H m_h \cos \left(\frac{2\pi h f_0 n}{f_S} + \phi_h \right) \right], \quad (1)$$

where $h \in 1..H$ are the harmonics, m_h is the modulation index at hf_0 , f_S is discrete signal’s sampling frequency and ϕ_h is an arbitrary phase shift assumed to be constant. We assume $a(n)$ to be non-negative. With purely sinusoidal amplitude modulation ($H = 1$), $a(n)$ is completely specified by the f_0 component, i.e., by m_1 and ϕ_1 . In natural voiced fricatives, the noise $w(n)$ is colored and the underlying modulation shape $a(n)$ is not a pure sinusoid.

B Amplitude modulation in fricatives

During voiced frication, transglottal pressure and laryngeal tension maintain phonation. Glottal vibration causes AM indirectly through variations in flow through the constriction, assumed to be fixed (Shadle, 1985). Although the presence of AM noise in voiced fricatives

is broadly acknowledged, fundamental questions remain. How does the result of glottal vibration reach the constriction? How does this perturbation affect the jet and the turbulence it forms? How does the perturbed turbulence go on to generate sound?

In his view of speech production, Stevens (1971) models noise sources under *static* aerodynamic conditions. From this perspective, the opening and closing of the vocal folds during one glottal cycle turns the flow on and off. Thus, the power of dipole sources is modulated proportional to V^6 while the flow velocity depends on the area and pressure across the constriction, ΔP_C (other sources accordingly). All these changes are considered to happen simultaneously. Noise source fluctuations up to 15 dB (i.e., $m \sim 0.7$) have been attributed to this mechanism (Stevens, 1971).

However, this static view belies the true complexity of the aeroacoustics. Experiments on mechanical models of the VT reveal considerable irregularity near the glottis which changes drastically in character moving downstream (Barney et al., 1999). Also, the phase of the noise modulation suggests that it takes time for the flow to generate noise (Jackson and Shadle, 2000).

An alternative view attributes AM to the interaction of the sound pressure wave created by phonation and the formation of turbulence in the jet exiting from the fricative constriction (Jackson and Shadle, 2000). Turbulent flows often exhibit large-scale regularity at certain ranges of flow rate and Reynolds number, $Re = \rho V D / \mu$, (Sinder, 1999); furthermore, unstable vortex formation is sensitive to acoustic interference and a sound wave near the jet's natural Strouhal number ($St = f_0 D / V$) regularizes or *forces* the turbulence, causing rotational flow structures to grow periodically (Simcox and Hoglund, 1971; Crow and Champagne, 1971). Forced modulation depths of $m \sim 0.2$ at $St=0.3$ were reported for their jet configuration. In voiced fricatives ($0.05 \leq St \leq 0.2$), voicing sets up the forcing wave, which then interacts with the jet to produce periodic vortices that convect downstream. These structures modulate the noise generation as, for example, the vortex train passes the obstacle. A further possible effect comes from the interaction of glottal vortices with the turbulence formed in the fricative jet, but data are incomplete (Barney et al., 1999).

There is little quantitative data published describing the acoustic characteristics of AM noise in fricatives, and it remains unclear exactly how those of the pulsed noise relate to the voicing that induces the pulsing, which this paper seeks to address. In a study of modulation phase, Jackson and Shadle (2000) gave some measure of average modulation depth in sustained voiced fricatives: their results range from 0 dB in the case of the bilabial fricative [β] to 2 dB in the case of [z] ($m \sim 0.25$); modulation for the other fricatives tended to cluster around 1 dB ($m \sim 0.1$).

In contrast to literature on modulation spectrograms (e.g., Tchorz and Kollmeier (2002)), signal processing techniques for cochlear implants (e.g., Rosen et al. (1999)) and temporal aspects of speech intelligibility (e.g., Shannon et al. (1995)), the modulation frequencies of interest in this paper are much higher ($>100\text{Hz}$). The primary object is to characterize the relationship of the modulation index at f_0 to other properties of speech, such as voicing strength, place of articulation and individual speaker differences. In particular, the modulation was examined at the voice fundamental frequency f_0 , plus harmonics $2f_0$ and $3f_0$. To quantify the observed modulation, and to move toward an understanding of the forcing mechanism, an estimate of the modulation index \hat{m} was computed from recorded speech signals.

C Noise models in speech synthesis

In simple models of voiced fricatives, the individual contributions from voicing and frication sources are summed to form the output: voicing as a volume-velocity source at the glottis; frication as a pressure source at the supraglottal constriction. Flanagan’s electrical analogue model was one of the first to incorporate AM of the fricative source (Flanagan and Cherry, 1969). Band-passed Gaussian noise (0.5–4 kHz) was multiplied by the squared volume velocity at the constriction exit, including the d.c. component. Sondhi and Schroeter (1987) employed a similar model for aspiration at the glottis, gated by a threshold Reynolds number; for frication they placed a volume-velocity source 0.5 cm downstream of the constriction exit (or at the lips for /f, v, θ, ð/). Klatt treated aspiration and frication identically, modulating

the noise source by a square wave (50% burst duration) that was switched on during voicing, to achieve the effect he wanted (Klatt, 1980).

In Scully’s work (Scully, 1990; Scully et al., 1992), noise generation was based on Stevens’ static experiments (Stevens, 1971): the strength of the pressure source was proportional to $\Delta P_C^{\frac{3}{2}}$, where ΔP_C is the pressure drop across the constriction. This source, depending on slowly-varying articulatory and aerodynamic parameters, was applied equally to aspiration and frication sources. Since ΔP_C across the supraglottal constriction was lower for voiced than voiceless fricatives, this equation yields the weaker frication source, but does not encode any modulation. However, motivated by perceptual test results, aspiration noise was modulated, using the rapidly-varying glottal area.

In his PhD thesis, Sinder (1999) presented a model for fricative production based on aeroacoustic theory. Once the necessary flow-separation conditions had been met, vortices were shed which convected along the tract, generating sound as they went, particularly when encountering an obstacle. In the Portuguese articulatory synthesizer of Teixeira et al. (2003), the volume velocity at the fricative constriction is based on the flow at the glottis and transfer functions computed for noise sources at several instants during an f_0 pitch period, allowing them to activate and deactivate.

D Organization

This research aims to extend current knowledge of AM noise generation by examining the relationship between the forcing glottal wave and modulation depth. Both sustained fricatives and fluent fricatives embedded in phrases were analyzed to provide modulation values suitable for integration into model-based speech synthesizers. Section 2 describes the speech recordings, the algorithm and parameters for estimating modulation depth in voiced frication. Section 3 discusses problems in applying the estimation algorithm to mixed-source speech and presents the preprocessing technique used to overcome them. The estimation procedure’s overall accuracy is then discussed. Section 4 presents results for the sustained and fluent fricatives, considering effects of gender, place of articulation (PoA), vowel con-

text and voicing strength. Section 5 draws together the main findings, relating them to earlier AM work, explanations of the forcing mechanism and current understanding of AM perception.

II METHOD

A Speech recordings

Two sets of fricative recordings were designed to capture the range of aeroacoustic conditions achievable by the human vocal apparatus in steady phonation and those typically realized in fluent speech. The *sustained fricative* set included laryngeal measurement of the vocal fold vibration and calibrated sound pressure from a microphone at a fixed distance from the subjects' lips for 16 subjects (12M, 4F). The *fluent-speech fricative* set provided a more natural environment with nonsense words framed in a standard phrase for 8 subjects (4M, 4F). Four subjects took part in both experiments. Numbers of male and female subjects reflected availability, while providing at least four of each sex, so gender effects are treated cautiously. Subjects were unpaid staff and students of the University of Surrey, age range 20–35, all with British RP accents.

1 *Sustained fricatives*

Fricatives [ʒ,z,ð,v] were spoken in isolation. Both male and female subjects was asked to produce two types of utterance at three pitch settings, $f_0 \in \{125, 150, 175 \text{ Hz}\}^1$. The first utterance type was an uninterrupted fricative where the subject smoothly adjusted loudness from their quietest possible fricative to loudest, and again to quiet, and loud (~ 3 s in total). The second type consisted of three separate sustained fricatives of increasing intensity (~ 1 s each). Each recording was preceded by a pitch-reference tone and short (2-s) pause to allow subjects to tune their pitch. Having three settings enabled an analysis for f_0 . In total 24 recordings were made for each speaker (4 fricatives \times 3 f_0 values \times 2 types).

Speech audio and electroglottograph (EGG) signals were captured simultaneously on PC

by a Creative Labs Audigy soundcard via a Sony SRP-V110 desk (2 channels at 44.1 kHz, 16 bit): mono audio from a Beyerdynamic M59 microphone, and EGG from a Laryngograph Lx Proc PCLX with adult-sized electrodes. A 1 kHz calibration tone was measured at the microphone and a B & K Type 2240 SPL meter, both 10 cm from the loudspeaker. During recording, subjects placed their head in a movement-restricting support and were instructed to keep still, to control the lip-microphone distance to within a few millimeters of 10 cm, at lip level and $\sim 45^\circ$ to the line of sight.²

2 *Fluent fricatives*

Speech-like tokens of $F = /f, \beta, s, z, \theta, \delta, f, v/$ were recorded from nonsense $/VF\emptyset/$ words with $V = /a, i, u/$, embedded in the phrase “What does $/VF\emptyset/$ mean?”, using an acoustically sheltered cubicle and equipment as above. Subjects were given two kinds of prompt: a randomized list of sentences and an audio recording of the list read by one author in time to a metronome (with pauses for response), played through single-ear headphone.³ To promote natural, fluent speech, subjects were left free to move their head. For each speaker, 216 sentences were recorded (9 reps. of every VF pair).

B Measuring modulation depth

1 *Estimating the modulation index m_h*

With modulated broadband noise, the carrier signal $w(n)$ is an unknown random variable, which can be modeled as Gaussian white noise, and the signal $x(n)$ is as in Eq. 1. To estimate m_h , the instantaneous magnitude of the signal is taken $|x(n)| = |w(n)|a(n)$ which, unlike the modulated noise signal, contains a periodic component at f_0 and its strength is proportional to m_1 . To extract this component, the Fourier transform is computed $\bar{X}(k) = \mathcal{F}\{|x(n)|\}$, applying a Hamming window and zero padding:

$$\bar{X}(k) = \mathcal{F}\{|w(n)|\} \otimes \left[\Delta(k) + \sum_{h=1}^H \frac{m_h}{2} (\Delta(k - hk_0) e^{j\phi_h}) + \sum_{h=1}^H \frac{m_h}{2} (\Delta(k + hk_0) e^{-j\phi_h}) \right], \quad (2)$$

where \otimes denotes convolution, $\Delta(\cdot)$ the Fourier transform of the window function, and $k_h = hNf_0/f_S$ is the frequency bin that contains harmonic hf_0 . Figure 3(d) shows the *modulation spectrum*, $\bar{X}(k)$, for frication noise from a [z] token modulated at $f_0 \approx 150$ Hz, where the spike occurs. Hence, modulation index m_h can be estimated by comparing the coefficients at hf_0 and d.c.: $\hat{m}_h = 2|\bar{X}(k_h)|/|\bar{X}(0)|$.

2 Allowing for pitch variation

Although the processing window is short enough to exclude major changes in fundamental frequency, pitch variation within a window smears the modulation energy at each harmonic. To compensate for variable pitch and spectral smearing from windowing, our estimate \hat{m}_h was based on the area under the spike at k_h and above the noise floor, including adjacent bins as appropriate⁴. This defined \tilde{k}_h as the contiguous set of bins under the k_h spike (see Fig. 3(d)). For the spike around zero frequency, a range of bins was also aggregated, $\tilde{0}$. Thus, with noise floor $\hat{\theta}^2 = \frac{1}{N} (1 - \frac{2}{\pi}) \sum_{k=0}^{N-1} |X(k)|^2$, a estimate similar to that above was formed:

$$\hat{m}_h = 2 \left(\frac{\sum_{\tilde{k}_h} |\bar{X}(k)|^2 - \hat{\theta}^2}{\sum_{\tilde{0}} |\bar{X}(k)|^2 - \hat{\theta}^2} \right)^{1/2}. \quad (3)$$

III APPLICATION TO SPEECH

A Periodic energy mixed with noise

Since voiced fricatives comprise periodic energy mixed with frication in the time waveform and much of the spectrum, it is not trivial to isolate the noise component for analysis. The f_0 component itself is confined to low frequencies (< 400 Hz) and can easily be removed by high-pass (HP) filtering without losing any significant amount of fricative noise. However,

bands of periodic energy, or voicing harmonics, persist into the higher spectral regions of the fricative noise, especially near formant frequencies.

For most speech sounds, interest is focused on the first two or three formants (up to perhaps 4 kHz) as higher formants tend to be weaker and are less important perceptually. In normal, fluent voiced fricative speech, voicing is often weak and its formants are rarely detectable much above 3 kHz; in strong fricatives with a loud voicing component (as in our sustained fricatives corpus), formants can be found up to 5 or 6 kHz. Consider the spectrum of a strongly voiced [v] in Figure 4. Fig. 4(b) shows the dominant periodic energy in the 0–4 kHz region (a harmonic spectrum with four defined formant peaks at 1.3, 2.2, 3.2 and 3.7 kHz); in Fig. 4(c), the spectrum is purely aperiodic, with no harmonics that can be ascertained in the 7–16 kHz range; in Fig. 4(a), 4–7 kHz contains mainly aperiodic energy though with a defined formant at 6.2 kHz, which can be seen in both the spectrum and the spectrogram in Fig. 4(d).

The effect on apparent modulation depth of mixing periodic energy with frication noise should be considered. Given that formants are damped resonances excited periodically by voicing at f_0 , they will tend to have a fluctuating envelope similar to that of the aperiodic component. Unless the peaks are in phase with the bursts of frication, the presence of voicing will attenuate the apparent modulation depth of the noise. Consider the fricative [v] in Figure 4. The spectrogram shows strongly modulated frication noise above 4 kHz, as well as fluctuating peaks in formant energy at lower frequencies. Careful inspection reveals that the pulses of frication are out of phase with the pulsed formant energy. Amplitude envelopes (or modulation signals, $a(n)$) for different frequency bands are shown underneath as Figs. 4(e) and (f), showing how they differ in phase. Fig. 4(e) compares amplitude fluctuation in the overwhelmingly periodic, 1–4 kHz band (thick line, cf. Fig. 4(b)), to the mainly aperiodic, 7–16 kHz band (thin line, cf. Fig. 4(c)). The phase difference between the two modulation signals is $\sim 170^\circ$.

Envelopes in Fig. 4(f) demonstrate the attenuation of apparent modulation from combining the out-of-phase periodic and aperiodic energy components. HP filtering with cutoff

$f_{\text{HP}} = 1$ kHz (thick solid line) removes the f_0 component *and first formant* but leaves the remaining periodic and aperiodic energy intact. Since the voicing source is stronger, the modulation signal is dominated by the periodic formant energy; the similarity in phase to the ‘periodic only’ 1–4 kHz band (thick solid line in Fig. 4(e)) confirms this. However, in comparison to the ‘periodic only’ case, the depth of modulation has been reduced; this is due to the out-of-phase aperiodic energy in the region 4–16 kHz. Raising f_{HP} to 3.5 kHz (dashed line) excludes most (but not all) of the periodic energy (see formant at 3.7 kHz in the spectrum and spectrogram), which evens out the periodic and aperiodic components. Modulation shape and depth are disrupted, and the phase of the modulating signal resembles neither of the previous cases. A further increment to $f_{\text{HP}} = 4$ kHz (thin line) excludes the last strong formant (a weaker one remains at ~ 6 kHz) and the resulting envelope is similar, if weaker, to the ‘aperiodic only’ 7–16 kHz band (thin line in Fig. 4(e)).

Since we are interested only in modulation of the frication noise, it is paramount that the *aperiodic component* is successfully isolated before applying Eq. 3 to estimate the modulation depth. As Figure 4 demonstrates, failure to remove periodic energy can seriously affect the accuracy of m_1 estimation for the frication noise. Periodic components could be removed by HP filtering with f_{HP} high enough to exclude all likely periodic energy.

Although fixing f_{HP} at a higher value has the advantage of effective removal of periodic energy, it substantially limits the bandwidth of noise from which modulation depth is measured. This causes two problems: first, modulation is unlikely to be uniform throughout the frication noise spectrum (see Sec. B); second, filtering AM noise removes some modulated sidebands which gives under-estimated modulation depth (see Sec. E).

B Non-uniformly modulated noise

Thus far, the noise signal has been treated as Gaussian white noise. In voiced fricatives, the carrier noise $w(n)$ is not white, but colored (filtered) depending on PoA. The spectral composition of the noise does not directly affect the modulation of different frequency regions. However, it cannot be assumed that the mechanism responsible for modulation in fricatives

produces uniform modulation across all frequencies; in fact, spectrograms of voiced fricatives suggest that noise in very high frequency regions (>8 kHz) is more modulated than in the main region (3–7 kHz). More work is needed to understand how the modulation mechanism produces uneven modulation depths across the noise spectrum.

Figure 5 shows a short portion (100 ms) of a strongly modulated [ʒ] that happens to lack strong voicing formants, allowing analysis of different frequency bands without interference from periodic energy. In the spectrogram, the frication noise looks modulated throughout the spectral range, but the weaker noise above 10 kHz comes in more distinct and separated bursts compared to the mid-range noise. This observation is borne out by analysis: amplitude envelopes for three spectral bands (magnitude signals, low-pass filtered at 700 Hz to catch the first few modulation harmonics) illustrate variations in the modulation signal through the noise spectrum. In the 3–6 kHz range (Fig. 5(b)), the modulation signal is noisy and its fundamental is weak ($m_1 = 0.56$). For 6–10 kHz (Fig. 5(c)), m_1 grows to 0.71, the waveform becomes more regular, and the periodic structure of the modulation signal emerges, with steep-sided, rather than sinusoidal, pulses. At 12–22 kHz (Fig. 5(d)), modulation at the fundamental is almost complete ($m_1 = 0.98$) and the waveform has regularized into a train of sharp (steep-sided) pulses separated by a noticeable gap. This is akin to the ‘fundamental saturating under the action of its harmonic’ described by Crow and Champagne (1971), where the fundamental can increase no further; instead, a significant harmonic structure develops where the modulation signal begins to adapt from sinusoid to pulse train. Thus, basing measurement of noise on the upper frequency bands could lead to an over-estimation of m_1 with regard to the full spectrum of frication noise. To balance the need for effective removal of periodic components and accurate estimation of modulation depth, the voiced fricative signals were preprocessed using a technique designed to segregate periodic and aperiodic energy.

C Pitch-scaled harmonic filtering

Separating periodic and aperiodic energy from a mixed-source signal is not a straightforward signal processing task. For speech signals, Yegnanarayana et al. (1998) and Jackson (2000) have proposed algorithms based on comb-filtering of harmonics using adaptive pitch data. By testing the algorithms on synthetic signals, and through informal listening tests, they have shown that speech can be effectively decomposed into periodic and aperiodic streams. In this study, Jackson (2000)’s decomposition algorithm, the *pitch-scaled harmonic filter* (PSHF, described in detail in Jackson and Shadle (2001)), was adopted as preprocessing to the modulation estimation procedure. Figure 6 shows the effect of applying the PSHF to 500 ms of [z] from the sustained fricative corpus. In spectrograms before and after, (a) and (c), the effects of pitch-scaled filtering are evident — formants below 4 kHz have been removed, although there remains some trace of the voicing fundamental. To ensure complete removal of the fundamental, high-pass filtering at a low frequency ($f_{\text{HP}} = 1$ kHz) was applied in addition to the PSHF. In Figs. 6(b) and (d), the effect on modulation depth of HP filtering employed alone is compared to the combination of PSHF and HP filtering. The HP filtered magnitude waveform, $|x(n)|$, from the PSHF’s aperiodic signal (Fig. 6(d)) shows deeper and sharper modulation. This was confirmed by measurements of m_1 which gave an increase of 0.11 (from 0.46 to 0.57) after application of the PSHF. The increase is attributed to the attenuating effect of periodic energy on modulation, described in Section A.

D Processing conditions

Choice of window size is a trade-off between modulation depth resolution and time resolution, which affects variability such as from pitch glides. Simulations using synthesized signals evaluated different window sizes (see Table 1). So, m_1 was estimated with a 100 ms window and a 5 ms step size for the sustained fricative corpus; for the fluent fricatives, a shorter 30 ms window was used. Processing windows were zero-padded to $N = 2^{15}$ points. The required values of f_0 were obtained from analysis of the EGG signal, when available; otherwise, from

the speech signal.

For later comparison, voicing strength v_1 was defined as the spectral amplitude at f_0 in the audio signal prior to high-pass filtering. For sustained fricatives, where subjects’ lip–microphone distance was strictly controlled and the microphone calibrated, v_1 is expressed as SPL (in Pa). For fluent-speech fricatives, the calibration to SPL was estimated by comparing RMS measurements averaged over all fluent-speech fricative waveforms to a calibrated test utterance recorded with the sustained fricatives. This estimated voicing strength \hat{v}_1 acts as a guide for comparing results from the two experiments.

E Evaluation of modulation estimates

In estimating the underlying modulation depth for a section of voiced frication, errors come from three sources. Error A is due to the nature of the noise signal: random variation inevitably gives to the impression of some small modulation component. Error B is introduced by the modulation estimation procedure (Sec. B), as a kind of bias. Finally, in the case of real voiced fricatives, imperfections in the preprocessing (Sec. C) will introduce further artifacts, error C . Simulation tests were conducted to evaluate the magnitude of the combined estimation error $A+B$. These tests involved making estimates of the modulation index from Gaussian white noise samples with an imposed amplitude modulation.

Summary results for two window sizes are given in Table 1 under three voicing conditions, incorporating descending pitch glides and random pitch variation, or jitter. Errors between true and estimated values are given in terms of average bias and variance, quoted as standard deviation. In all cases, the bias was small compared to the deviation, which was twice as high for the short (23 ms) window as for the longer (93 ms) window. The longer window gave errors of ± 0.04 (2σ) on the estimates under typical speech conditions.

Establishing the magnitude of error C is less simple. Filtering partially fills in ‘valleys’ in the temporal waveform and thus reduces in modulation depth. Eddins (1993) ran simulations to evaluate the effect of band-pass filtering on m_1 of modulated white noise varying the bandwidth, $f_{\text{BW}} \in \{0.2, 0.4, 0.8, 1.6\}$ kHz. He concluded that modulation depth was

‘relatively unaffected’ for these filter conditions. Our own simulations investigating the effects of limiting bandwidth of modulated noise by high-pass filtering showed the effect to be secondary, increasing the range to ± 0.05 at the highest 11-kHz cut-on frequency (lowest bandwidth). The 1-kHz HP filter applied here has negligible effect, as does the erroneous removal of some noise by the PSHF.

To validate use of the PSHF on voiced fricatives, its effect with known modulation was assessed. Phonetically-trained subjects recorded voiced and noise components of voiced fricatives separately by producing sustained *voiceless* fricatives, introducing phonation, then gradually relaxing the constriction, leaving just voicing.

Recordings were edited to give voicing plus frication noise with an imposed m . Random, 100 ms sections of frication with known m (0.1–1) were mixed with sections of voicing (from same speaker/fricative) with amplitude varying 0–15 dB in comparison to the frication (periodic to aperiodic ratio, PAR) and preprocessed (Sec. C) before measurement of m .

PAR significantly affected the accuracy of estimation for each preprocessing stage. For strongly voiced fricatives, the error with HP filtering was much improved by applying the PSHF. Where the voicing component was insignificant, HP filtering produced a better estimate alone, due to PSHF artifacts.

In 1000-trial simulation, where PAR varied freely (as in natural fricatives), overall bias was 0.03, suggesting a tendency to overestimate, and 2σ range rose to 0.10 (cf. 0.18 with HP filtering only). While justifying the use of the PSHF, this result is misleading in some respects. Most voiced fricatives are not very strongly voiced, so estimates produced using only HP filtering are fairly reliable; hence accuracy increases only slightly with the PSHF. Tokens with strong voicing, where using the PSHF gave large increases in accuracy, were less common but characteristic of particular speakers or PoAs. Without the PSHF, results for those speakers and fricatives would be inaccurate, though a fraction of all fricatives. Thus, the PSHF improves comparability of results.

IV MODULATION RESULTS

A The \hat{m}_1 vs. v_1 relationship

Figure 7 summarizes \hat{m}_1 for all the data. To explore the relationship between voicing strength, v_1 or \hat{v}_1 , and modulation depth \hat{m}_1 , v_1 ranges spanning all the data (0–0.3 Pa SPL for sustained fricatives; 0–0.07 Pa SPL for fluent-speech fricatives) were split into equal bins (0.01 and 0.003 Pa bin width for sustained and fluent-speech respectively). The \hat{m}_1 vs. v_1 relationship is represented as voicing strength bin centers plotted against average \hat{m}_1 reading for that bin. Histograms show number of frames in each bin.

In producing the sustained fricatives, very high or low levels of voicing were seldom used, resulting in an approximately normal distribution. Voicing levels in the fluent-speech case were significantly lower, as expected for short, intervocalic fricatives. The skew of the distribution toward lower values of \hat{v}_1 in Fig. 7(d) can be attributed to voice dynamics in intervocalic voiced fricatives: voicing rapidly decreases in amplitude as frication begins and either remains low until the vowel onset, or ceases (devoicing) (Pincas, 2004).

Figure 8 shows the low voicing strengths ($0 \leq v_1$ or $\hat{v}_1 \leq 0.05$). There are fewer data frames for the fluent-speech fricatives as each was so short; at higher values of \hat{v}_1 where \hat{m}_1 was stronger, the lack of data leads to wide error intervals, compared with the sustained fricatives. The \hat{m}_1 vs. \hat{v}_1 curve for sustained fricatives levels off sharply at $v_1 = 0.03$ Pa, where modulation saturates, $\hat{m}_1 \approx 0.5$. Above $v_1 = 0.04$ Pa, \hat{m}_1 remains constant until $v_1 = 0.25$ Pa (Fig. 7(a)), where the data become too sparse to give meaningful results. For fluent-speech fricatives, \hat{m}_1 saturated earlier, by $\hat{v}_1 = 0.02$ Pa, and was slightly lower (~ 0.4) than for sustained fricatives. Above $\hat{v}_1 = 0.03$ Pa, data was sparse (Fig. 8(b), histogram counts fall below 250) and the bin averages beyond $\hat{v}_1 = 0.05$ Pa should be interpreted with caution.

Figure 9 (thick lines) illustrates the \hat{m}_1 vs. v_1 relationship for individual speakers. In sustained fricatives, saturation occurred at a similar point (0.03–0.04 Pa) for all subjects except MD; saturation values of \hat{m}_1 were also similar for each speaker; quoted \hat{m}_1 readings

were at 0.055 Pa, from the bin following saturation. Although mean \hat{m}_1 ranged 0.13–0.64, the distribution ($\mu = 0.43$, $\sigma = 0.12$) shows that, on average, speakers’ modulation tends to lie around the 0.4–0.5 mark.

Given the imbalance of male to female subjects, only cautious comment can be made in comparison of their results. No difference is immediately discernible in \hat{m}_1 at saturation, although statistical comparison reveals a slight difference in mean and distribution (male $\mu = 0.40$, $\sigma = 0.12$; female $\mu = 0.50$, $\sigma = 0.05$).

Individual differences in degree of modulation could correspond to an aspect of voice quality. Significantly, the limiting values of \hat{m}_1 came well before modulation was complete ($m_1 = 1$), and imply saturation of a physical AM mechanism.

For the four speakers who took part in both experiments (JP, PJ, AT and RG; two male, two female), comparison of results suggests similar behavior across experiments (except JP, whose patterns for sustained and fluent-speech fricatives are obviously different). The fluent-speech curves for subjects PJ and RG appear to match the initial portions of their respective sustained fricative curves well. AT’s fluent-speech and sustained fricative data complement one another, providing reliable readings at lower voicing strengths and a continuing pattern at higher strengths respectively.

B Effect of place of articulation

Differences among the four English voiced fricatives are seen in Figure 10. Error intervals are wider than those in Figures 7 and 8 but the basic \hat{m}_1 vs. v_1 relationship remains the same for all four fricatives, with varying saturation parameters for each PoA. The curve for [z] (thick solid line) stands out: it is the quickest to saturate (at $v_1 \approx 0.035$) and does so at a highest modulation depth. Furthermore, the transition from the rising, linear part of the curve to the saturated part is more abrupt than for other fricatives. The high modulation depth at saturation for [z] in Fig. 10 is common to most speakers: 14 of 16 subjects have [z] as the most heavily modulated fricative at $v_1 = 0.05$ Pa.⁵ These findings echo previous results for [z] in fluent speech (Pincas and Jackson, 2004). Considering the alternative views of modulated

noise production discussed in the Introduction, there are several possible interpretations. According to the static view, the constriction area, A_C , determines the pressure drop across the constriction, ΔP_C , relative to that at the glottis (Stevens, 1971). So, for [z], which has a marginally smaller constriction (0.17 cm^2) compared to other places (0.19 cm^2) (Narayanan et al., 1995), the modulation of ΔP_C , and hence of the flow velocity and noise intensity, would be lesser ($m \sim 0.6$). However, area differences may not be the most significant factor. The monopole, quadrupole and dipole sources for each PoA have varied amplitudes and critical Reynolds numbers due to their particular geometry, which could account for the observed differences in m .

The view based on forced turbulence has the advantage that the greater acoustic pressure fluctuation in the smaller constriction would strengthen forcing, tending to raise noise modulation. Yet the precise geometry could have a more substantial influence, for the reasons above, but also since the constriction-obstacle distance and Strouhal number are critical for this mechanism. Modulation is maximal 2–6 diameters from the jet exit, i.e., 1–3 cm, and forcing closer to the natural Strouhal number can double the modulation (Crow and Champagne, 1971). Furthermore, the distribution of sources (e.g., dipoles along the upper lip in non-sibilants [v,dh]) affects modulation phase ϕ_h through turbulence convection Coker et al. (1996). Thus distributed sources exhibit reduced modulation. Note that alveolar fricatives have the most concentrated dipole source at the lower incisors.

C Harmonic structure of $a(n)$

The aeroacoustic processes that produce AM noise in voiced fricatives might be thought of as follows: a forcing glottal wave, $d(n)$, interacts with a noise generation process to produce AM noise near the fricative constriction. Following reflections within the VT, the noise radiates as the voiced fricative signal, $x(n)=a(n)w(n)$. The shape of $x(n)$'s envelope is described by the modulating signal $a(n)$ applied to an unmodulated frication noise signal $w(n)$ and its modulation spectrum has a component m_1 at the fundamental. In relating $d(n)$ to $a(n)$, the results discount the linear hypothesis that $d(n)$ is proportional to $a(n)$ (i.e., that the

underlying modulation is identical in shape to the glottal wave that initiated it). This is demonstrated by the saturation of \hat{m}_1 , the fundamental component of $a(n)$, as a function of v_1 , the fundamental component of $d(n)$. Yet, the full $d(n)$ to $a(n)$ mapping requires further clarification.

Observations confirm that even the most strongly modulated frication noise shows negligible components above the second harmonic (i.e., only m_1 and m_2 are significant) and in many cases m_2 is so weak as to blend into the background fluctuations, leaving m_1 only. This is true even when the forcing wave shows significant harmonic structure. Figure 3 gives an example of such a situation for a token of [z] taken from the corpus: the forcing wave $d(n)$ is represented by the low-pass filtered audio waveform. This is compared to the high-pass filtered magnitude waveform $|x(n)|$, whose spectrum has peaks at harmonics of the modulating signal $a(n)$. Note how the harmonic structure of $d(n)$ in Fig. 3(b) was not preserved in the modulation spectrum of the noise, shown in Fig. 3(d).

Figure 11 shows \hat{m}_h values at the first and second harmonics using the familiar binning procedure. As v_1 increases, a significant modulation harmonic \hat{m}_2 does arise and \hat{m}_3 was detectable. Although the results cannot rule out the possibility that m_2 was caused by the same harmonic in the forcing wave (i.e., v_2), it seems more likely that they conform to the behavior observed by Crow and Champagne in a comparable study using turbulent jets forced by a *pure sinusoid* from a loudspeaker (Crow and Champagne, 1971).

Figure 9 shows the harmonic analysis for individual subjects. Some speakers (cf., JPLM and MZ-RG) show relatively little modulation at the higher harmonics. To ascertain whether this difference depends on the forcing wave’s harmonics (voice quality variation), or on natural variation in the modulating signal, requires further investigation.

D Effect of f_0

Figure 12 analyzes the effect of voicing pitch on modulation depth for male and female subjects for both experiments. The relationship between voicing strength, v_1 or \hat{v}_1 , and modulation depth \hat{m}_1 is plotted in Figs. 12(a,b,d,e) grouped by fundamental frequency f_0

(bin edges determined by dividing the range of 95% of the data into three equal-width bins). The measured distributions of f_0 are shown in Figs. 12(c) and (f).

Figure 12(c) reveals that subjects were not very successful in attaining the required f_0 (125, 150, 175 Hz), in the sustained fricatives experiment. Female subjects, as might be expected, had particular difficulty with the lower pitches. The distribution of f_0 data is thus wider than anticipated, but nevertheless provides an appropriate base for analysis. In the fluent-speech fricative experiment, where subjects spoke at their natural pitch, f_0 distributions are significantly tighter. As a result, data are sparse in the lower pitch bins from female subjects (150–180 Hz and 180–210; Fig. 12(e)), and dominated by one subject at higher voicing strengths, producing an anomalous curve (KC in Fig. 9, bottom right).

Fundamental frequency of voice has little consequence for the relationship between voicing strength and modulation depth, with similar shaped curves throughout. Furthermore, there is some suggestion in the sustained fricative experiment that male subjects (Fig. 12(a)) produce higher modulation at lower f_0 for all but one voicing level. However, this pattern is not replicated in any other results and we conclude that f_0 is not an important influence on modulation depth.

E Perceptual Considerations

The combination of harmonic and amplitude-modulated noise sources is special to voiced frication and presents an interesting and complex picture from a psychoacoustic perspective. On a basic level, it is known that modulation effects a change in the quality of the noise component, creating a sensation of ‘roughness’ (Zwicker and Fastl, 1999). However, most previous work on the perception of AM noise is limited in relevance to voiced fricatives, due to their short duration and the presence of voicing.

1 *Detection of amplitude modulation*

The extent of the percept created by AM depends, of course, on the depth of noise modulation (m), but also on a number of other factors. Numerous authors have reported the relationship

between the detection threshold of AM noise θ with sinusoidal envelope and its frequency f , referred to as the Temporal Modulation Transfer Function, or TMTF (Bacon and Viemeister, 1985; Patterson et al., 1978; Viemeister, 1979). Detection thresholds at each f are measured using a forced choice paradigm: subjects must differentiate the modulated stimulus interval from one or two accompanying unmodulated noise intervals. The modulation depth of the target interval is adjusted gradually according to the subject’s responses, to yield finally an estimate of the detection threshold. Thresholds are low in the region of frequencies applicable to speech (e.g., $\theta \approx 0.13$ at $f = 125$ Hz), although they increase with f by ~ 3 – 4 dB/octave; hence a small difference in detection threshold is expected for typical male and female voices. The TMTF also has implications for the detectability of modulation at harmonics of f_0 . With $f_0 = 125$ Hz, the second harmonic’s modulation detection threshold is $\theta \approx 0.18$ ($f = 250$ Hz). Given that m_2 tends to be below this level (Fig. 9), modulation at harmonics above f_0 is not likely to be detectable. In addition, deeper modulation at f_0 could mask shallower modulation at $2f_0$ in an effect in the modulation domain akin to regular psychoacoustic masking in the frequency domain (see literature on ‘modulation masking’, e.g., Houtgast (1989)).

Stimulus duration also affects our ability to detect AM. In the literature, thresholds are almost always based on 500 ms stimuli, yet voiced fricatives are much shorter.⁶ Lee and Bacon (1997) investigated the effect of stimulus duration on modulation detection threshold and showed that shorter stimuli did indeed yield higher thresholds.

The added effect of voicing is extremely hard to predict. A low-frequency voicing component significantly louder than the noise component, as with non-sibilants [ð] and [v], would produce masking (in the regular, frequency-domain sense (Fletcher, 1940); i.e., an increase in absolute detection threshold of the noise). The consequences of this decrease in audibility for the detection of AM are not known, but it may be of note that Viemeister (1979) found minimal difference in AM detection for stimuli presented at different levels.

The combination of tone and noise further complicates the detection of AM. Wakefield and Viemeister (1985) performed what appears to be the only investigation into AM noise

detection in the presence of a pure sinusoid with f equal to that of the modulating signal. Their results suggest a key role for the phase between tone and modulation, with the possibility of detection being *enhanced* where the two are in phase. The finding is hard to generalize, however, since they used only 3 kHz bandwidth noise.

2 *Perceptual coherence*

Correlated temporal patterns across disparate spectral components (such as those present in modulated noise) create or reinforce distinct auditory ‘objects’ (Bregman, 1990). A commonly cited example of this effect is comodulation masking release, or CMR (for a review see Verhey et al. (2003)), where modulation imposed on a masking noise band causes the detection threshold of a tone at the band’s center to *decrease* (detection improves) as the bandwidth of the noise is widened, even past the critical bandwidth (CB) (Hall et al., 1984); this contrasts with the classical psychophysical masking paradigm where increasing noise bandwidth beyond the CB has no effect on masking Fletcher (1940), and suggests that listeners are able to use the ‘comodulated’ temporal pattern of the noise to improve stream segregation and detection of the tone. The relevance of stream segregation and modulation to speech has previously been demonstrated by Hermes (1991), who found that the cohesiveness of synthesized breathy vowels was enhanced by modulation of the aperiodic component. If this effect extends to voiced fricatives, modulation of frication could enhance the integrity or intelligibility of speech in noise.

V CONCLUSION

In voiced fricatives, phonation induces amplitude modulation of frication noise. A technique was developed to estimate the depth of modulation and applied to turbulence noise from sustained and fluent-speech fricatives. Modulation depth rose approximately linearly with voicing strength for low voicing levels (below ~ 63 dB SPL); it saturated at a similar voicing level for different fricatives and speakers, although its value at this point varied. For example,

modulation depth at a voicing strength of 0.04 Pa SPL (immediately after saturation) was largest for [z] (0.65; cf. 0.44 for [ʒ], 0.37 for [ð], 0.34 for [v]). Previous perceptual studies of modulated noise suggest that the levels of modulation observed are detectable. Further work could establish how amplitude-modulated noise in fricatives serves as a phonetic cue or voice-quality characteristic, and investigate the aeroacoustic mechanism responsible for producing modulation.

Notes

¹Since the effect of f_0 on m was unknown, this control ensured comparability of results, especially between male and female speakers.

² A short lip-microphone distance helped to capture quiet frication over any background or electric noise.

³ A pilot experiment revealed that some subjects had difficulty keeping their place on the printed list while speaking. The audio prompting was designed to aid them, but also as a natural control on speech rate and intonation.

⁴ The present method diverges here from that used in Pincas and Jackson (2004).

⁵In contrast, saturation points and levels for the remaining fricatives, whilst relatively similar and consistently distinct from [z], vary for each speaker with no clear pattern. This could be explained by articulatory configurations varying less across speakers for [z], but more for the other fricatives which tend either to cause difficulty (e.g., [ʒ] is quite rare in English) or to be produced in a variety of ways (e.g., [ð] varies in degree of tongue protrusion). The slightly narrower confidence intervals for [z] at higher voicing strengths concur.

⁶Mean intervocalic fricative durations, averaged over 216 repetitions by 8 subjects, according to Pincas (2004): [v]–70 ms, [ð]–69 ms, [z]–92 ms and [ʒ]–101 ms. ANOVA shows significant ($p < 0.0005$) difference between sibilants [z,ʒ] and non sibilants [ð,v] but no significant difference within these pairs.

References

- Bacon, S. and N. Viemeister (1985). Temporal modulation transfer functions in normal-hearing and hearing-impaired subjects. *Audiology* 24, 117–134.
- Barney, A., C. H. Shadle, and P. Davies (1999). Fluid flow in a dynamic mechanical model of the vocal folds and tract. 1. measurements and theory. *J. Acoust. Soc. Am.* 105(1), 444–455.
- Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organisation of Sound*. MIT, Cambridge, MA.
- Coker, C. H., M. H. Krane, B. Y. Reis, and R. A. Kubli (1996). Search for unexplored effects in speech production. In *Proc. Int. Conf. Spoken Language Processing 1996*, Philadelphia, PA, Volume 14(6), pp. 415–422.
- Crow, S. C. and F. H. Champagne (1971). Orderly structure in jet turbulence. *J. Fluid Mech.* 48, 547–591.
- Eddins, D. (1993). Amplitude modulation detection of narrow-band noise: Effects of absolute bandwidth and frequency region. *J. Acoust. Soc. Am.* 93(1), 470–479.
- Flanagan, J. L. and L. Cherry (1969). Excitation of vocal-tract synthesizers. *J. Acoust. Soc. Am.* 45(3), 764–769.
- Fletcher, H. (1940). Auditory patterns. *Rev. Mod. Phys.* 12, 47–65.
- Hall, J., M. Haggard, and M. Fernandes (1984). Detection in noise by spectro-temporal pattern analysis. *J. Acoust. Soc. Am.* 76, 50–56.
- Heid, S. and S. Hawkins (1999). Synthesizing systematic variation at the boundaries between vowels and obstruents. In *Proc. ICPHs*, San Fransisco, pp. 511–514.
- Hermes, D. J. (1991). Synthesis of breathy vowels: some research methods. *Speech Comm.* 10(5-6), 497–502.

- Houtgast, T. (1989). Frequency selectivity in amplitude-modulation detection. *J. Acoust. Soc. Am.* 85(6), 1676–1680.
- Jackson, P. J. B. (2000). *Characterisation of plosive, fricative and aspiration components in speech production*. Ph. D. thesis, Dept. Electronics and Computer Science, University of Southampton.
- Jackson, P. J. B. and C. H. Shadle (2000). Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. *J. Acoust. Soc. Am.* 108(4), 1421–1434.
- Jackson, P. J. B. and C. H. Shadle (2001). Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. *IEEE Trans. on Speech & Audio Proc.* 9(7), 713–726.
- Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* 67(3), 971–995.
- Lee, J. and S. P. Bacon (1997). Amplitude modulation depth discrimination of a sinusoidal carrier: Effect of stimulus duration. *J. Acoust. Soc. Am.* 101(6), 3688–3693.
- Lighthill, M. (1952). On sound generated aerodynamically. I. General theory. In *Proceedings of the Royal Society*, Volume 211, pp. 564–587.
- Lighthill, M. (1954). On sound generated aerodynamically. II. Turbulence as a source of sound. In *Proceedings of the Royal Society*, Volume 222, pp. 1–34.
- Lofqvist, A., T. Baer, N.S. McGarr, and R. Seider-Story (1989). The cricothyroid muscle in voicing control. *J. Acoust. Soc. Am.* 85(3), 1314–1321.
- Lofqvist, A., L. L. Koenig, and R. S. McGowan (1995). Vocal tract aerodynamics in /aCa/ utterances: Measurements. *Speech Comm.* 16, 50–66.
- Narayanan, S. S. and A. A. Alwan (2000). Noise source models for fricative consonants. *IEEE Trans. on Speech & Audio Proc.* 8(2), 328–344.

- Narayanan, S. S., A. A. Alwan, and K. Haker (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *J. Acoust. Soc. Am.* 98(3), 1325–1347.
- Pastel, L. (1987). Turbulent noise sources in vocal tract models. Master’s thesis, MIT, Cambridge, MA.
- Patterson, R., D. Johnson-Davies, and R. Milroy (1978). Amplitude-modulated noise: The detection of modulation versus the detection of modulation rate. *J. Acoust. Soc. Am.* 63(6), 1904–1911.
- Pincas, J. (2004). The interaction of voicing and frication sources in speech: An acoustic study. Master’s thesis, School of Electronics and Physical Sciences, University of Surrey.
- Pincas, J. and P. J. B. Jackson (2004). Acoustic correlates of voicing-frication interaction in fricatives. In *Proc. From Sound to Sense*, Cambridge, MA, pp. C73–C78.
- Rosen, S., A. Faulkner, and L. Wilson (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *J. Acoust. Soc. Am.* 106(6), 3629–3636.
- Scully, C. (1990). Articulatory synthesis. In W. J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 151–186. Dordrecht, Netherlands: 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.
- Shadle, C. (1990). ‘Articulatory-Acoustic Relationships in Fricative Consonants’ in *Speech Production and Speech Modelling*, W.J.Hardcastle and A. Marchal (eds.), pp. 187–209. Kluwer Academic Publishers.
- Shadle, C. H. (1985, March). The acoustics of fricative consonants. Technical Report 506, RLE, Massachusetts Institute of Technology.

- Shadle, C. H. (1995). Modelling the noise source in voiced fricatives. In *Proc. 15th Int. Congress on Acoustics* Trondheim, Norway, Volume 3.
- Shannon, R. V., F.-G. Zen, V. Kamath, J. Wygonski, and M. Ekelid (1995). Speech recognition with temporal cues. *Science* 270, 303–304.
- Simcox, C. D. and R. F. Hoglund (1971). Acoustic interactions with turbulent jets. *Trans. Am. Soc. Mech. Eng. J. Bas. Eng.* 93(1), 42–46.
- Sinder, D. J. (1999). *Speech synthesis using an aeroacoustic fricative model*. Ph. D. thesis, Dept. Electrical Engineering, Rutgers University, New Brunswick, NJ.
- Sondhi, M. M. and J. Schroeter (1987). A hybrid time-frequency domain articulatory speech synthesiser. *IEEE Trans. on Acoust., Speech & Sig. Proc.* 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.
- Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA 02142-1493, USA: The MIT Press.
- Tchorz, J. and B. Kollmeier (2002). Estimation of the signal-to-noise ratio with amplitude modulation spectrograms. *Speech Comm.* 38, 1–17.
- Teixeira, A., L. M. T. Jesus, and R. Martinez (2003). Adding fricatives to the portuguese articulatory synthesiser. In *Proc. Eurospeech 2003*, Geneva, Switzerland, pp. 2949–2952.
- Verhey, J. L., D. Pressnitzer, and I. M. Winter (2003). The psychophysics and physiology of comodulation masking release. *Exp. Brain Res.* 153, 405–417.
- Viemeister, N. (1979). Temporal modulation transfer functions based upon modulation thresholds. *J. Acoust. Soc. Am.* 66(5), 1364–1380.
- Wakefield, G. H. and N. F. Viemeister (1985). Temporal interactions between pure tones and amplitude modulated noise. *J. Acoust. Soc. Am.* 77(4), 1535–1542.

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

Zwicker, E. and H. Fastl (1999). *Psychoacoustics: Facts and Models, 2nd Edition*. Springer-Verlag, Berlin.

Table 1: Estimation errors (bias, deviation) over all frames in 100 files versus analysis window size, with $8\times$ zero padding. Values are averaged across modulation index $m \in \{0.0, 0.1, \dots, 1.0\}$.

f_0 (Hz)	Jitter (%)	Window size	
		1024 (23 ms)	4096 (93 ms)
150	0.0	-0.004, 0.037	0.003, 0.020
160–140	0.5	-0.005, 0.037	0.006, 0.019
180–120	1.5	-0.017, 0.039	0.003, 0.020

Figure 1: Sound production mechanisms in schematic mid-sagittal view of the vocal tract in voiced fricative configuration: (G)lottis, (C)onstriction, (O)bstacle, (L)ip termination. Acoustic sources: \triangle periodic, \circ monopole, \diamond quadrupole, and ∞ dipole noise.

Figure 2: (a) Spectrogram, (b) Waveform, and (c) Pitch track of /VF/ transition in [a:ʒ] token. 16 kHz bandwidth.

Figure 3: Illustration of the harmonic structure of the voicing signal (top row) and the modulating signal (bottom row) for 100 ms of [z] ($f_0 \approx 150$ Hz). (a) Audio waveform low-pass filtered at 1 kHz. (b) Audio spectrum up to 500 Hz. (c) Magnitude of waveform high-pass filtered at 9 kHz. (d) Modulation spectrum. Dashed lines in spectra indicate noise floor.

Figure 4: (a) LPC spectrum (order 40), (b) Close-up of spectrum in region 0–4 kHz, (c) Close-up of spectrum in region 7–16 kHz, (d) Spectrogram (5 ms, Hanning window, $4\times$ zero-padded, fixed gray-scale, frequency-aligned with LPC spectrum and time-aligned with amplitude envelopes), and (e,f) Amplitude envelopes (magnitude signal, low-pass filtered at 200 Hz) for 50 ms section of sustained [v] ($f_0 \approx 153$ Hz, $f_s = 32$ kHz). Individual amplitude envelopes are for different frequency bands, f_{BP} . (e) $1 \leq f_{BP} \leq 4$ kHz (thick line, periodic energy) and $7 \leq f_{BP} \leq 16$ kHz (thin line, aperiodic energy); dashed horizontal lines on spectrogram identify these frequency regions. (f) $1 \leq f_{BP} \leq 16$ kHz (thick line, mainly periodic), $3.5 \leq f_{BP} \leq 16$ kHz (dashed line, balanced mix of periodic and aperiodic) and $4 \leq f_{BP} \leq 16$ kHz (thin line, mainly aperiodic).

Figure 5: (a) Spectrogram, and (b,c,d) Time-aligned waveforms (light gray) with amplitude envelopes (black lines, magnitude signal low-pass filtered at 700 Hz) for 100 ms section of sustained [ʒ] ($f_0 \approx 152$ Hz, $f_S = 44.1$ kHz). Individual amplitude envelopes are for different frequency bands, f_{BP} , with axes scaled to $\pm 2 \times$ RMS amplitude (indicated by dashed lines; notice the different scale for each band). (a) $3 \leq f_{BP} \leq 6$ kHz; (b) $6 \leq f_{BP} \leq 10$ kHz; (c) $12 \leq f_{BP} \leq 22$ kHz. \hat{m}_1 values estimated for individual frequency bands as in Sec. II.B.

Figure 6: 500 ms section of [z]; $f_0 \approx 125$ Hz. Left column: before PSHF. Right column: after PSHF. (a,c) Fixed gray-scale spectrograms. (b,d) $f_{HP} = 500$ Hz filtered magnitude waveforms, $|x(n)|$, for 300–400 ms portion of signal; \hat{m} estimates obtained as in Sec. II.B.

Figure 7: Top: Modulation depth \hat{m}_1 as a function of voicing strength v_1 or \hat{v}_1 . Bottom: v_1 or \hat{v}_1 distribution histograms for sustained fricatives (left column) and fluent-speech fricatives (right column). Data are means and counts of values falling within ± 0.01 Pa bins (sustained fricatives) or ± 0.003 Pa bins (fluent-speech fricatives). Error bars show standard error.

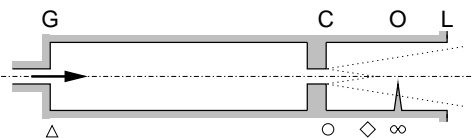
Figure 8: (a) Modulation depth \hat{m}_1 as a function of voicing strength v_1 or \hat{v}_1 , and (b) v_1 or \hat{v}_1 distribution histogram for sustained fricatives (thick line) and fluent-speech fricatives (thin line). Data are means and counts of values falling within ± 0.003 Pa bins. Error bars show standard error.

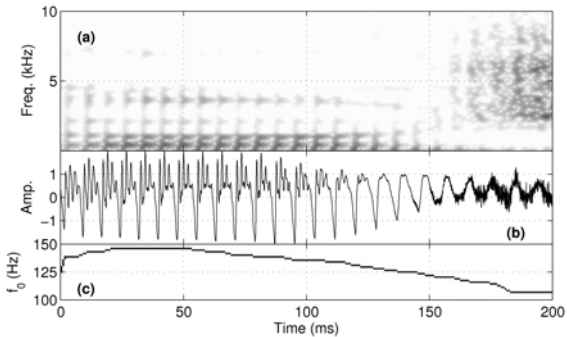
Figure 9: Modulation depths at the fundamental frequency \hat{m}_1 (thick line), second harmonic \hat{m}_2 (thin line) and third harmonic \hat{m}_3 (dashed line), versus voicing strength v_1 or \hat{v}_1 for individual speakers for sustained fricatives (top four rows) and fluent-speech fricatives (bottom two rows). Data are means and counts of values falling within ± 0.005 Pa bins. Error bars show standard error. Subjects' initials with male/female indication are given. m_1 values quoted for sustained fricatives are mean \hat{m}_1 over the voicing strength bin $0.05 \leq v_1 < 0.06$ Pa.

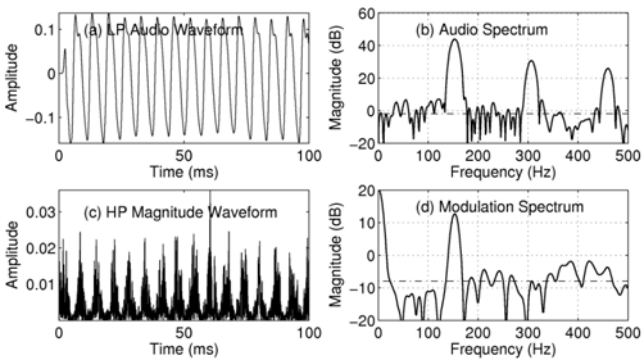
Figure 10: Modulation depth \hat{m}_1 as a function of voicing strength v_1 or \hat{v}_1 for (a) sustained, and (b) fluent-speech fricatives: $[\delta]$ – solid thin; $[v]$ – dotted thin; $[z]$ – solid thick and $[\mathfrak{z}]$ – dotted thick. Data are means and counts of values falling within ± 0.005 Pa bins (sustained fricatives) or ± 0.003 Pa bins (fluent-speech fricatives). Error bars show standard error.

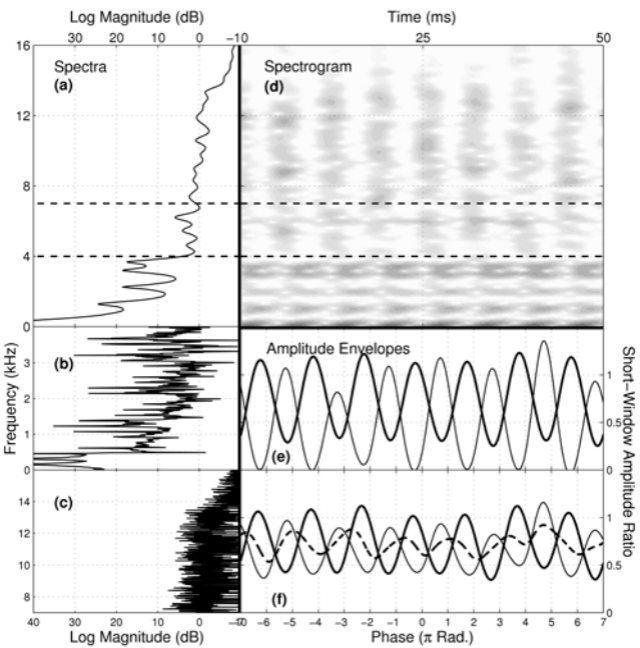
Figure 11: Modulation depths at the fundamental frequency \hat{m}_1 , second harmonic \hat{m}_2 and third harmonic \hat{m}_3 versus voicing strength v_1 or \hat{v}_1 for (a) sustained fricatives, and (b) fluent-speech fricatives. Means from all tokens. Data are means and counts of values falling within ± 0.003 Pa bins. Error bars show standard error.

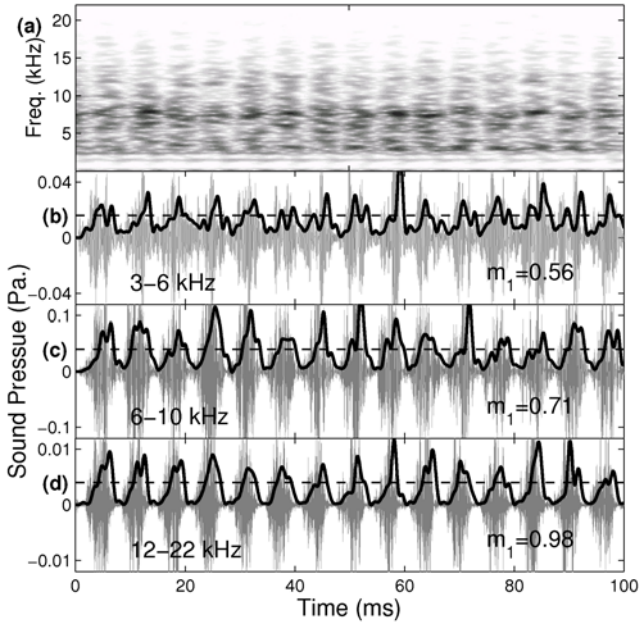
Figure 12: Top: Modulation depth \hat{m}_1 as a function of voicing strength v_1 or \hat{v}_1 for (a) sustained fricatives, male subjects; (b) sustained fricatives, female subjects; (d) fluent-speech fricatives, male subjects; (e) fluent-speech fricatives, female subjects. f_0 data divided into 3 equally-spaced pitch bins (different for each plot). In general: low range (thin line), middle range (medium line), and high range (thick line). For specific bin values see legends. Data for each f_0 bin are means of all frames whose measured f_0 falls into that bin. Voicing strength, v_1 or \hat{v}_1 , binning used ± 0.005 Pa bins. Error bars show standard error. Bottom: measured f_0 distribution histograms for (c) sustained fricatives, and (f) fluent-speech fricatives. Data are means and counts of values falling within ± 20 Hz bins from all tokens for male (gray bars) and female (clear bars) speakers.

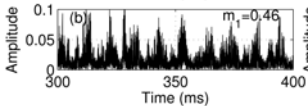
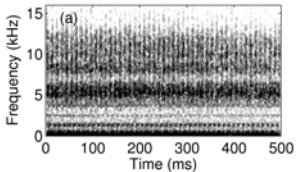
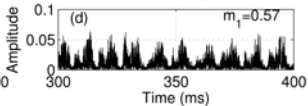
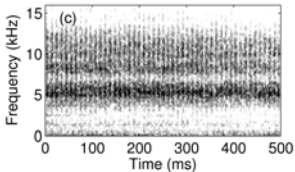






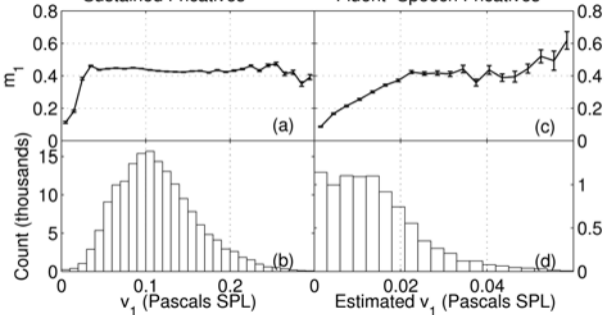


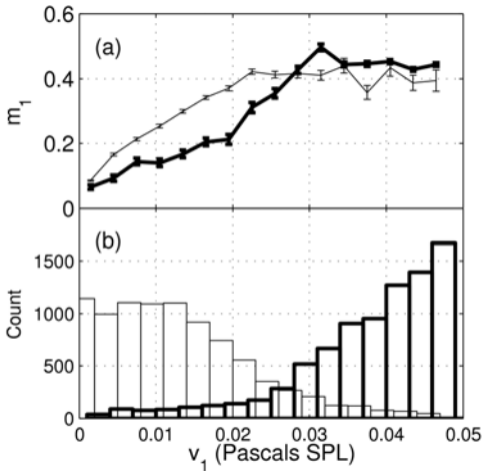


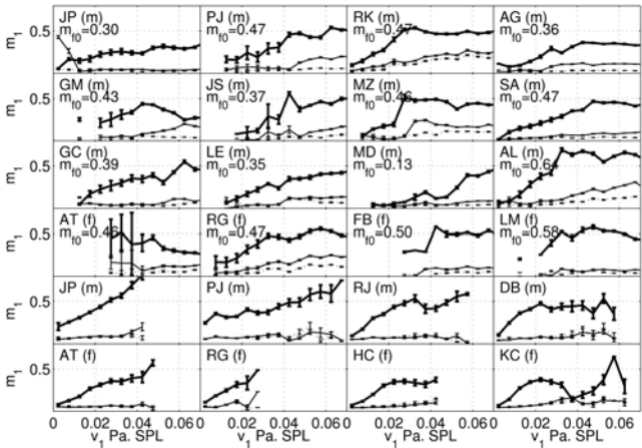
Before PSHF**After PSHF**

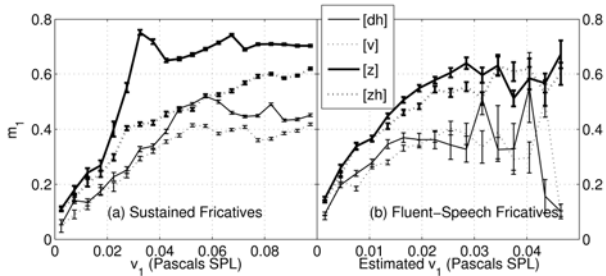
Sustained Fricatives

Fluent-Speech Fricatives

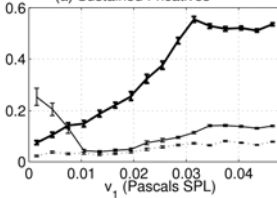




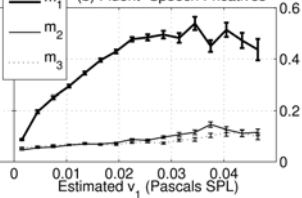




(a) Sustained Fricatives



(b) Fluent-Speech Fricatives



Sustained Fricatives

Fluent-Speech Fricatives

