Amplitude Modulation of Frication Noise by Voicing Saturates

Jonathan Pincas and Philip J. B. Jackson

Centre for Vision, Speech & Signal Processing, Univ. of Surrey, Guildford, Surrey GU2 7XH, UK. [j.pincas,p.jackson@surrey.ac.uk]

Abstract

The two distinct sound sources comprising voiced frication, voicing and friction, interact. One effect is that the periodic source of the noise source of up to 15 dB have been attributed to this cause by the periodic interruptions of glottal vibration are possible interpretation, then, is that the variations in airflow modulated frication is normally assumed to result from peri-

Fricatives are speech sounds produced by forcing air through a narrow constriction superior to the glottis, generating turbulence noise within the jet itself, or at/along a physical obstacle further downstream. English has voiceless and voiced fricatives at four places of articulation: labiodental, dental, alveolar and postalveolar, giving a total of eight fricative phonemes.

Voiced fricatives are generally distinguished by the presence of glottal and fricative sources, and this mixed excitation lends them their ‘buzzy’ quality. The characteristics of voiced frication do not arise simply from the linear combination of its component sources. The articulatory, aerodynamic and acoustical conditions required by and resulting from the simultaneous production of glottal vibration and frication noise raise the possibility of ‘mutual interaction effects’ [1]: the presence of each source causes the other to be changed in character from the case where it occurs in isolation. The focus of this paper, amplitude modulation (AM) of the frication component, is one such effect; others include mutual amplitude reduction [2], changes in fundamental frequency of voicing [3], and spectral changes in the voicing component (before, during and after frication) [4] and in the frication-noise component [5].

Although the presence of AM noise in voiced fricatives is widely acknowledged, the underlying mechanism is still not fully understood. During voiced frication, transglottal pressure and laryngeal tension conditions combine to maintain glottal vibration. The results of phonation are twofold and travel through the vocal tract at different speeds [6]: a jet of air leaving the glottis generates sound via pressure fluctuation, and sets up hydrodynamic motion (mean flow velocity). Amplitude-modulated frication is normally assumed to result from periodically pulsed flow through a fixed-area constriction [5]. A possible interpretation, then, is that the variations in airflow caused by the periodic interruptions of glottal vibration are responsible for modulation. Indeed, fluctuations in the amplitude of the noise source of up to 15 dB have been attributed to this mechanism [2]. The aerodynamic situation, however, is not so straightforward. Mechanical model studies have shown that airflow conditions along the vocal tract are complex, with the high degree of periodic airflow fluctuation immediately superior to the glottis being largely eroded further up the vocal tract [7]. Although it is thus hard to predict the patterns of airflow at any potential constriction in the vocal tract, it is unlikely that this mechanism is solely, or even partly, responsible for modulation. It is, in fact, more likely that AM is attributable to the interaction of the pressure wave created by phonation and the turbulent jet formation process at the fricative constriction [6].

Periodic, large-scale regularity in unstable flows is a common phenomenon in fluid mechanics. Although it has been shown that jets issuing from circular constrictions can exhibit large-scale regularity at increasing Reynolds numbers without any accompanying sound-pressure field [8], it appears that when a pressure wave at or near the natural Strouhal number of the jet is introduced, the cyclical fluctuation in the jet flow is significantly boosted. This is possible because unstable jet formation is sensitive to the presence of acoustic waves [8], which regularise, or force, turbulence generation.

For voiced fricatives, the forcing pressure wave is that set up by glottal vibration. This wave then interacts with the jet formation process at the supraglottal constriction, producing a periodically-fluctuating jet. In the majority of cases, the creation of amplitude-modulated noise probably depends heavily on the jet striking a downstream obstacle, such as the lips or teeth. Phase differences between the glottal and noise-modulation signals reported by Jackson and Shadle support this interpretation [6].

There has been little quantitative study devoted to the acoustic characteristics of AM noise in fricatives. For fricatives embedded in fluent speech nonsense words, Pincas and Jackson found that modulation depth tracked voicing strength quite closely and that the voiced fricative [z] was generally more heavily modulated than others [1]. Jackson and Shadle also published limited data relating to amplitude of modulation in various voiced fricatives [6]: their results range from 0 dB in the case of [8] to 2 dB in the case of [z]; modulation for the other fricatives tended to cluster around 1 dB.

This study aims to extend our current knowledge of the AM noise generation process by exploring the relationship between the forcing glottal wave and modulation depth. The data obtained is also apt for integration into a speech synthesis system.

2. Method

2.1. Speech Data Acquisition

The Corpus: Sustained English voiced fricatives ([v,ð,ʒ,ʃ]) were produced by 16 speakers (12 M, 4 F). Each fricative was produced separately with voicing at 125, 150 and 175 Hz. Each fricative-pitch combination was preceded by a calibration tone.
played through a loudspeaker and a short (2-s) pause allowing the subject to attain the correct voicing pitch. Two repetitions of each combination were performed. The first was an uninterrupted fricative where the subject smoothly adjusted loudness from the quietest fricative they could produce to relatively loud, and back again (~3 s in total). The second repetition consisted of three separate sustained fricative bursts with gradually increasing amplitude, each lasting approximately 1 s. For each speaker 24 recordings were made (4 frics × 3 pitches × 2 reps).

**Recording:** 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 with 16-bit resolution): mono audio from a Beyer-dynamic M59 microphone, and EGG from a Laryngograph Lx Proc PCLX with adult-sized microphones. The microphone was calibrated by comparing a 1 kHz tone played through a loudspeaker at 10 cm to an SPL measurement made with a Bruel and Kjær Type 2240 SPL meter at the same distance. Subjects placed their head in a support to minimise movement throughout recording and the calibrated microphone was placed 10 cm away, at lip level and at approximately 45° to the subject’s line of sight. The EGG signal provided accurate pitch information which was used by the modulation depth estimation algorithm.

### 2.2. Measuring Modulation Depth

**Fundamentals of AM:** Modulation depth, \( m \), is most often given in standard index form, which can be conceptualised as the fraction of the carrier signal that the modulated signal varies by, e.g., if \( m = 0.5 \), then the signal fluctuates by 50% above and below its original, unmodulated value. In most applications of AM (such as in acoustics or telecommunications), \( m \) ranges from 0 (unmodulated) to 1 (completely modulated).

In AM, the amplitude of the carrier signal, \( w(n) \), is modified by a modulating signal, \( a(n) \), to produce an amplitude-modulated signal, \( x(n) = w(n)a(n) \). In the case of a periodic modulating signal, \( a(n) \) takes the form of a fundamental sinusoid of frequency \( f_0 \) plus its harmonics. Thus, we have

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

where \( h \in 1 \ldots H \). \( H \) are the harmonics, \( m_h \), is the modulation index at \( h f_0, f_s \) is the sampling frequency and \( \phi_h \) is an arbitrary phase shift which we assume to be constant. With purely sinusoidal amplitude modulation (\( H = 1 \)), the signal \( a(n) \) is completely specified by the \( f_0 \) component, i.e., by \( m_1 \) and \( \phi_1 \). In natural voiced fricatives, the underlying modulation shape is unlikely to be purely sinusoidal. Here, however, we will mainly be concerned with modulation at \( f_0 \), and so for ease of reference we will refer to \( m_1 \) as \( m_{f0} \). Where we refer to higher modulation harmonics, they are designated \( m_{2 f0}, m_{3 f0} \) etc.

**Estimating \( m_{f0} \):** In the case of modulated broadband noise, the carrier signal \( w(n) \) takes the form of a random variable which we model as Gaussian white noise and the signal \( x(n) \) is fully specified by Equation 1. To estimate \( m_{f0} \) we first take the instantaneous magnitude of the signal: \( |x(n)| = |w(n)|a(n) \), which contains a periodic component at \( f_0 \), the strength of which is directly proportional to \( m_{f0} \). Hence we compute its Fourier transform, \( \tilde{X}(k) = \mathcal{F} \{ |x(n)| \} \), first applying a Hamming window and zero-padding to \( N \) points (2^16):

\[
\tilde{X}(k) = \mathcal{F} \{ |w(n)| \} \otimes \left[ \delta(0) + \sum_{h=1}^{H} m_h \frac{1}{2} \left( \delta(\pm h k_0) e^{\pm j\phi_h} \right) \right]
\]

where \( \otimes \) denotes convolution, \( \delta(\cdot) \) the Dirac delta function, and \( k_0 = N f_0 / f_s \) is the frequency bin that contains \( f_0 \). Figure 1 presents a synthetic example of the modulation spectrum, \( \tilde{X}(k) \), for white noise modulated at \( f_0 = 150 \) Hz, where the spike occurs.\(^1\) Finally, \( m_{f0} \) can be calculated by comparing the magnitudes of the Fourier coefficients at d.c. and \( f_0 \):

\[
m_{f0} = 2 \frac{|\tilde{X}(k_0)|}{|\tilde{X}(0)|},
\]

where the factor of two leads to an estimate of the standard modulation index. Clearly, where \( m_{2 f0}, m_{3 f0} \) etc. are required in place of \( m_{f0}, k_0 \) is replaced by the relevant integer multiple.

**Isolating the frication noise:** In the case of voiced fricatives, the carrier noise \( w(n) \) would not be white, but coloured (filtered) depending on the fricative place of articulation. A further complicating factor is the presence of low-frequency voicing and excited formants mixed with the frication noise. Given that periodically-excited formants are damped oscillations pulsed at \( f_0 \), the presence of periodic energy normally serves to attenuate aperiodic modulation depth, unless the pulses are perfectly in phase with the bursts of frication noise. Since we are interested only in modulation of the frication noise, it is paramount that we successfully isolate the aperiodic component before applying the procedure outlined above. Efficient removal of periodic components is achieved by high-pass (HP) filtering with a cutoff frequency, \( f_{HP} \). However, since HP filtering also removes noise components below \( f_{HP} \), we would effectively only be measuring modulation for frication noise above \( f_{HP} \). Inspection of spectrograms suggests that modulation is unlikely to be uniform across the spectrum: noise in high-frequency regions looks to be more modulated than in lower regions, where it is more concentrated. Thus, biasing measurement to noise in the upper frequency bands will lead to an overestimation of modulation depth with regard to the full spectrum of frication noise, which is our ultimate object of interest. To balance the need for effective removal of periodic components and accurate estimation of modulation depth, we experimented with a 40th-order HP filter at six cutoff frequencies, \( f_{HP} \in \{ 0.7, 1.4, 2.7, 4.5, 8.4 \) and 11.5 kHz \}.

**Variable pitch:** Although the processing window employed was short enough to exclude major changes in fundamental frequency, pitch variation within a window would lead to modulation energy being spread around \( f_0 \). To compensate for variable pitch, we based our estimate \( \tilde{m}_{f0} \) on the area of the spike at \( k_0 \) (see Fig. 1). Upper and lower extremes of the base of the spike, \( k_{UL} \) and \( k_{LL} \) respectively, and hence its width, are dictated by the noise floor, \( m_{f0} = 0.5 \). Dashed line indicates noise floor.

\[^1\]Modulation does not alter the flatness of a white noise spectrum.


3. Results and Discussion

The $\tilde{m}_{f0}/v_{f0}$ relationship: In Fig. 2, readings for all speakers, fricatives, pitch levels and repetitions are combined. Results are presented for all HP cutoff frequencies. To explore the relationship between modulation depth and voicing strength, the $v_{f0}$ range 0.01–0.1 Pa SPL (up to 74 dB SPL) was split into 10 equally-spaced bins and readings within each bin averaged.

We begin by considering $\tilde{m}_{f0}$, the modulation depth at the fundamental, depicted by the black and blue lines in Fig. 2. Although data for very low voicing strengths was sparser, all bins have 95% confidence intervals narrower than 0.05 (modulation index), which is similar to predicted estimation error. The relationship between $\tilde{m}_{f0}$ and $v_{f0}$ is non-linear for all values of $f_{HP}$, with saturation occurring at approximately $v_{f0}=0.04$ Pa (at 10 cm) for all but the $f_{HP}=11.5$ kHz case. At the point of saturation, modulation index varies between approximately 0.5 (for $f_{HP}=700$ Hz) and 0.7 (for $f_{HP}=11.5$ Hz). At lower voicing strengths the curve rises almost linearly, with an increase in modulation index of between 0.12 (for $f_{HP}=700$ Hz) and 0.18 (for $f_{HP}=11.5$ Hz) for every 0.01 Pa increase in $v_{f0}$.

Effect for HP-cutoff frequency: The curves for different values of $f_{HP}$ appear to support the initial observations mentioned in Sec. 2.2. At a HP filter cutoff of 700 Hz, we expect a certain amount of voicing energy mixed with frication noise and thus probable underestimation of true modulation; Fig. 2 (thin black line) bears this out. Raising the cutoff to 1.4 kHz (dashed black line) eliminates most periodic energy without excluding a significant amount of friction noise and this modulation depth is better estimated. Notice how raising the cutoff further to 2.7 kHz (thick black line) produces little difference: periodic energy is already mostly eliminated and the bulk of the frication energy, for most places of articulation, remains above the cutoff frequency. However, raising $f_{HP}$ further to 4.5 kHz (thin blue line) and 8.4 kHz (dashed blue line) produces overestimation as measurement is biased to the more deeply modulated noise in the higher-frequency region. Raising the cutoff from 8.4 to 11.5 kHz (thick blue line) has little effect as most of the concentrated noise has already been excluded. We conclude, then, that HP filtering at approximately 1.5–3 kHz is suitable as pre-processing to the estimation procedure described in Sec. 2.2 and hence these results best reflect real modulation depths for voiced fricatives.

Harmonic structure of $a(n)$: The aerodynamic 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 vocal tract, the noise radiates as the voiced fricative signal, $x(n)$. The shape of $x(n)$’s envelope is described by the underlying modulating signal $a(n)$, which has a component $m_{f0}$ at the fundamental. In relating $d(n)$ to $a(n)$, note that the results discount the hypothesis that $d(n)$ is equal to $a(n)$ (i.e., that the underlying modulation is identical in shape to the forcing wave that initiated it). This is manifested by the saturation in the relationship between the fundamental components of $d(n)$ and $a(n)$: $v_{f0}$ and $m_{f0}$ respectively. Yet, the full $d(n)$ to $a(n)$ mapping requires further clarification.

Our observations confirm that even the most strongly modulated frication noise shows no detectable components above the second harmonic (i.e., only a fundamental and second harmonic are present) and in many cases the harmonic is so weak as to blend into the background fluctuations, leaving a fundamental 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. Notice how the harmonic structure of $d(n)$ (top right) is not preserved in the modulation of the noise (bottom right).
Along with the fact that the modulated noise in voiced fricatives will be perceived due to complicating factors such as their short length and presence of the low-frequency voicing signal (pseudosinusoidal from a loudspeaker [8]. In a comparable study using turbulent jets forced by a pure sinusoid, the similarly-shaped curves obtained by Crow and Champagne [3] were less perceptible due to the presence of the low-frequency noise. The difference in level of modulation is detectable. Further work could establish how amplitude-modulated noise in fricatives serves as a phonetic cue or voice-quality characteristic.

5. References