Number:

Problem	Points	Score
1 (a)	10	
1 (b)	10	
1 (c)	10	
1 (d)	10	
2 (a)	10	
2 (b)	10	
2 (c)	10	
3 (a)	10	
3 (b)	10	
3 (c)	10	
Total	100	

Notes:

1. The exam is closed books and notes. You are allowed one 8 1/2” x 11” double-sided sheet of notes.

2. Please indicate clearly your answer to the problem by some form of highlighting (underlining).

3. Your solutions must be legible and easy to follow. If I can’t read it or understand it, it is wrong. Random scribbling will not receive credit.

4. Please show ALL work. Answers with no supporting explanations or work will be given no credit.

5. Several problems on this exam are fairly open-ended. Since the evaluation of your answers is obviously a subjective process, we will use a market place strategy in determining the grade. Papers will be rank-ordered in terms of the quality of the solutions, and grades distributed accordingly.

1. Deep sea divers breath a mixture of air and helium called Heliox to avoid several problems associated with breathing compressed air under water. Heliox is lighter than air — its density is 75% lower than air. Speech produced while breathing Heliox sounds distorted (a classroom demonstration is provided).

(a) Predict the effect breathing Heliox has on the formant frequencies. Justify this answer using our linear acoustics model.

Recall from our acoustic tube model, resonances are at $f = c/4l$, where c is the velocity. Velocity is inversely proportional to the density. To be precise, for an ideal gas, $c = \sqrt{\frac{\gamma RT}{\sigma}}$, where γ is the adiabatic constant, R is the gas constant (J/mol K), T is the absolute temperature (K), and σ is the density (kg/mol). Hence, the formants (and velocity) will increase by a factor of $\sqrt{4/3}$. In reality, the velocity of sound in helium is about three times that of air, and about 1.5 to 2.5 times that of air in the helium/air mixtures that divers use.

Note that in grading this problem, I was primarily looking for an equation relating the formants to the velocity. Since we didn't cover details beyond that in class, I didn't expect the analysis above.

(b) Does the excitation signal change? Explain.

The excitation signal is created by stretching the vocal cords over the air pushed from the lungs. The frequency of vibration is related to the tension on the vocal cords, which is independent of the medium. Hence, the excitation largely stays the same, which implies key features such as fundamental frequency and voicing don't change.

c) For what value of the density (relative to air) would the speech become unintelligible to a human listener?

The example I played in class was barely intelligible. Suppose the speaker was breathing an even lighter gas. The formants would shift even higher, making them further outside of the range of formant frequencies humans expect. As the formants shift higher, our ability to recognize the sounds as speech diminishes. In the limit, if the formants shifted about the half sample frequency, we would have real problems. If the density were in the range of 8/3 or 16/3, the formants would shift to a point where the speech would be unintelligible (this shift would be perceived as 50% greater than the amount of shift demonstrated in class, which would push it beyond the limit of human perception).

(d) Design a system that would descramble the diver's speech and produce a normal sounding speech signal. Is such a system physically realizable?

Since only the formants shift, we need a way of scaling the smooth spectral structure, which represents the vocal tract) but keeping the excitation fixed. If we were to use the frequency scaling property of the Fourier Transform, we could shift the entire spectrum. This would be a good first-order approximation, and could be realized using FFTs and inverse FFTs. However, a better way to do this would be to use an advanced signal processing algorithm to separate the excitation from the vocal tract (such as linear prediction, cepstral analysis, or other such techniques which we have subsequently discussed), and then scale only the vocal tract model. Such a system is also physically realizable — see the URLs provided in class for more details.

2. Consider a language which has a phoneme set that only contains four English consonants: *b, p, d, t.*

(a) Describe the similarities and differences between these sounds in as much linguistic detail as possible.

p:	bilabial voiceless stop	b:	bilabial voiced stop
t:	lingua-alveolar voiceless stop	d:	lingua-alveolar voiced stop

The only difference between p and b is voicing. Similarly, the only difference between t and d is voicing. The only difference between p and t, and well as b and d, is the place of articulation. These sounds have very similar spectral signatures and are hard to distinguish. They are also hard for our foreign students to pronounce :)

(b) Do the assumptions we make to justify frame-based processing in speech recognition of spoken English hold for this type of signal? Explain.

If you only had to discriminate between four consonants, you would certainly not worry to a large extent about vocal tract resonances, periodicity, etc. Remember, the Fourier Transform spectrum is only valid for stationary signals (e.g., steady-state, deterministic, white noise). You would probably look at temporal features such as energy, the rate of change of energy, etc. You would also probably want to increase your frame duration and amount of overlap so you had a good chance to observe on frame in which the dynamic nature of the sound was clearly captured within the central portion of the window.

On the other hand, as we will see, our current feature sets are capable of recognized such sounds because they measure rate of change as well as absolute spectral behavior.

I accepted both sides of the argument on this question as long as your position was justified.

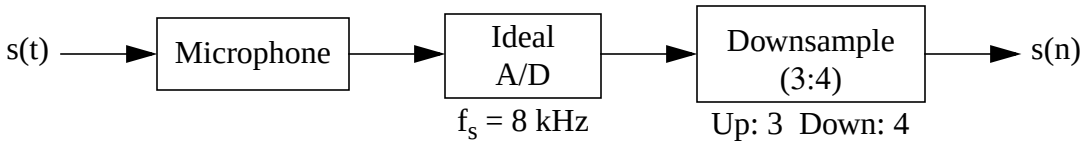
(c) Consider a voiced sound in this language produced with a fundamental frequency of 100 Hz. Would a listener perceive the 5th and 6th harmonics to be closer in frequency than the 20th and 21st harmonics? Explain.

Use our equation for mel: $mel = 2595\log 10(1 + f/700)$. Convert these frequencies to the mel scale: 151, 283, 1521, 1562 mel respectively. Take the difference:

$$(|f_1 - f_2| = 132) > (|f_{20} - f_{21}| = 41)$$

Hence, the first two harmonics would be perceived to be more distant in frequency. This makes sense since hearing is logarithmically sensitive to frequency. This is an example of why the bins in our filter bank design get wider with increasing frequency. We do not need as much resolution in the upper frequency range.

3. Consider the system shown below:



(a) Suppose the microphone can be modeled by this equation: $y(t) = \alpha x(t) + \beta x^2(t)$. Typically, $\alpha > \beta$. Sketch the spectrum of the output signal, s(n), over the frequency range [0, 8 kHz] for an input that consists of a sinewave at 1 kHz.

For the given input, the output is:

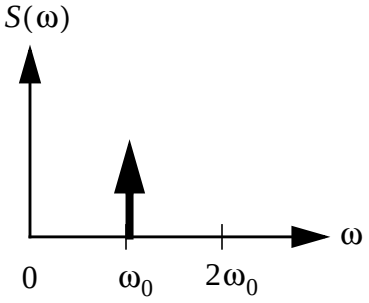
$$y(t) = \alpha A \sin(\omega_0 t) + \beta A^2 \sin^2(\omega_0 t) = \alpha A \sin(\omega_0 t) + \beta (A^2/2)[1 - \cos 2\omega_0 t]$$

The Fourier transform of this signal is:

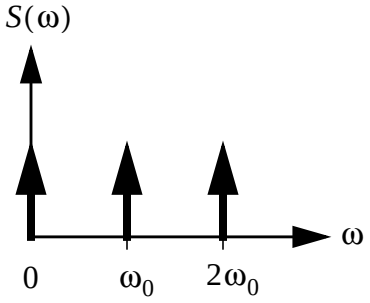
$$\begin{aligned} Y(\omega) = {} & 2\pi\beta(A^2/2)\delta(\omega) \\ & + \alpha A(-j\pi)[\delta(\omega - \omega_0) - \delta(\omega + \omega_0)] \\ & + \beta(A^2/2)(\pi)[\delta(\omega - 2\omega_0) + \delta(\omega + 2\omega_0)] \end{aligned}$$

Note that sampling at 8 kHz doesn't introduce aliasing. Further, downsampling doesn't change the frequency of the signal. Below is a sketch of the frequency response:

Input:



Output:



Note that the output is a digital spectrum, so it is periodic in the frequency domain. Its period is given by the new sample frequency: $(3/4)8 \text{ kHz} = 6 \text{ kHz}$. **Also, the output contains frequencies that the input did not contain due to the nonlinearity in the microphone.** This is why nonlinearities in a microphone are a bad thing.

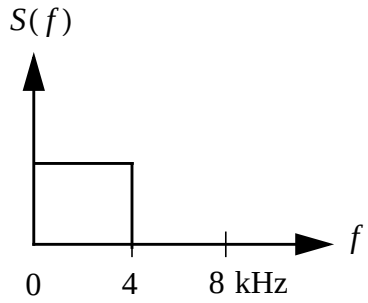
(b) Sketch the spectrum of the output signal for a white noise input (flat spectrum). Assume both the A/D and the downsampler use ideal low pass filters.

Let us define the input noise signal as $s(t) = Gw(t)$. Following the logic above, the input to the A/D will have a spectrum that can be described as follows:

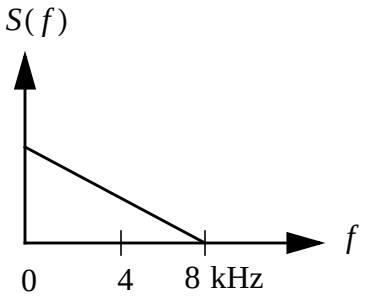
$$Y(\omega) = GW(\omega) + F[G^2 w^2(t)]$$

Multiplication in the time domain is equal to convolution in the frequency domain. Hence, the noise spectrum will be the convolution of its spectrum with itself:

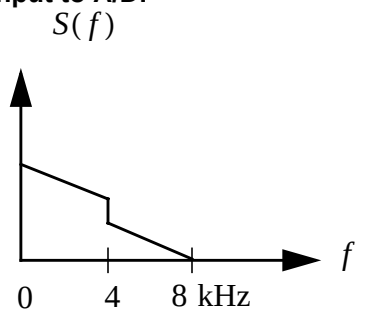
Input:



Squared:

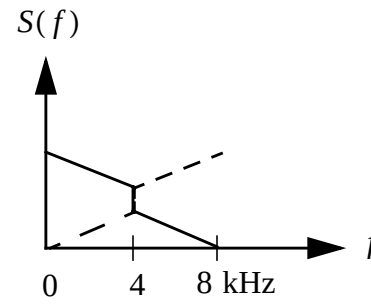


Input to A/D:

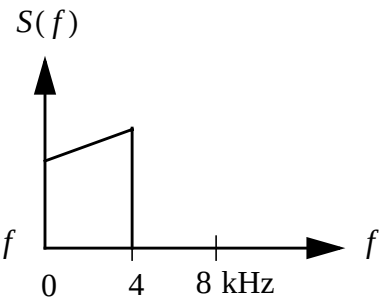


Next, the A/D will introduce aliasing, shown below. Following that, the downsampling will have to low-pass filter the spectrum at 3 kHz before downsampling. This will result in the spectrum shown below labeled output:

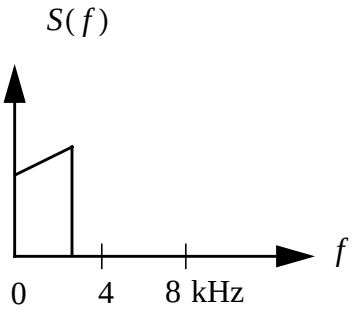
Aliasing:



Aliased Sampled Spectrum:



Downsampled Output:



The point is that the nonlinearities of the microphone again cause significant "in-band" distortion — this time due to aliasing. I was very lenient in grading this problem and did not expect this level of detail.

(c) Suppose the input signal is the sum of a 1 kHz and 1.5 kHz sinewave. Sketch the spectrum of the output signal. Explain the influence of the microphone on this result. What aspect of the microphone would you improve? Why?

The nonlinearity is going to create energy at frequencies other than the two original frequencies:

$$\begin{aligned} x^2(t) &= (A\sin(\omega_1 t) + B\sin(\omega_2 t))^2 \\ &= A^2\sin^2(\omega_1 t) + 2AB\sin(\omega_1 t)\sin(\omega_2 t) + B^2\sin^2(\omega_2 t) \\ &= A^2\left[\frac{1}{2} + \frac{1}{2}\cos(2\omega_1 t)\right] + B^2\left[\frac{1}{2} + \frac{1}{2}\cos(2\omega_2 t)\right] \\ &\quad + 2AB\left[\frac{1}{2}\cos((\omega_1 - \omega_2)t) + \frac{1}{2}\cos((\omega_1 + \omega_2)t)\right] \end{aligned}$$

As you can see, the presence of the terms $2\omega_1$, $2\omega_2$, $\omega_1 - \omega_2$, and $\omega_1 + \omega_2$ are particularly troublesome since these frequencies were not present in the original signal. We obviously need to reduce the nonlinearity in the microphone: $\alpha \gg \beta$.