Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

MODELING THE PERCEPTION OF PITCH-RATE AMPLITUDE MODULATION IN NOISE

Brian P. Strope[1,2] and Abeer A. Alwan[1]

[1]*Department of Electrical Engineering*
UCLA, 405 Hilgard Ave., Los Angeles, CA 90095, USA
[2]*Nuance Communications*
1380 Willow Road, Menlo Park, CA 94025, USA

1. Introduction

The robustness of speech communication depends on a structural hierarchy that is deeply embedded with redundancy. This structure exists in our language — sentences, phrases, words, syllables, phonemes — as well as in the rapidly varying acoustic details that cue these perceptions. Together, the stages of this hierarchy form a web of partially orthogonal dimensions. In typically noisy situations the listener is unlikely to perceive each of these representational units in explicit detail. Instead, partially corrupted cues are readily filled-in with expectations derived from other stages in the hierarchy: the listener may miss the phonetic segment but still reconstruct the word (or miss the word, but understand the gist of the phrase, etc.).

This chapter demonstrates that a similar redundancy exists for the detection of voicing in noise. Specifically, after power spectrum cues are removed, listeners can use amplitude modulation cues to detect voicing at low signal-to-noise ratios.

Because of this redundancy, amplitude modulation cues in voiced speech provide a salient, robust sensation of pitch that may be instrumental in recognizing speech in noise. In the current study, three psychoacoustic models are used to predict the temporal modulation transfer function (TMTF) and the detection of voicing for high-pass filtered naturally spoken fricatives in noise. Computational models based on waveform-envelope statistics and modulation filtering properties predict the TMTF data with a high degree of precision, and models derived from a summary autocorrelogram representation fit both the TMTF and high-pass filtered data sets.

1.1 Voicing in Speech Analysis and Automatic Recognition

During voiced speech, the vibration of the vocal folds excites time-varying resonances of the vocal tract. Given a sequence of feature vectors representing log-magnitude, spectral estimates of the vocal-tract transfer function across time, most automatic speech recognition (ASR) systems use a hierarchy of non-stationary stochastic models operating at progressively longer intervals of speech analysis (10–30 ms) and statistical modeling (at the representational level of the phonetic segment, word, phrase and sentence) to ascertain what was most likely to have been said [17]. However, ASR systems rarely use pitch or voicing information in this process.

Instead, the signal processing for feature vector extraction usually reflects some form of deconvolution, attempting to shield vocal-tract transfer-function estimates from the impact of the driving function. Linear prediction, for example, is used with a predictor polynomial that is significantly shorter than the anticipated glottal periodicity. Similarly, when homomorphic analysis is used for ASR, the high-quefrency cepstral terms (which can represent the periodic ripple across the spectral estimate resulting from a harmonic driving function), are ignored. Using Mel-frequency cepstral coefficients (MFCC), the initial spectral estimate is first averaged (in time) over multiple pitch periods and then integrated across frequency, providing an approximation of auditory frequency selectivity. The output is then logarithmically compressed and the discrete cosine transform is used to partially decorrelate the log-magnitude spectral estimate across frequency. Higher-order terms in the resulting cepstral vector are ignored. Integrating across time and frequency reduces the variance of the spectral estimate, and together with the truncated cepstral vector, nearly eliminates periodic source information.

Deconvolution is an important step for isolating the phonetic information about "*what was said*," from aspects of the prosodic information pertaining to "*how it was said*." But as the first processing stage it may be eliminating large parts of the perceptually salient information used by humans to identify and recognize speech in noisy environments.

Speech communication has evolved to be robust in noise. Redundancies are, therefore, ubiquitous. Perceiving speech under noisy conditions requires an intelligent use of the potentially unreliable, albeit redundant, multi-dimensional cues spread over wide-ranging time scales. While deconvolution must occur somewhere in the recognition process, blindly eliminating a potential wealth of redundant cues may not be appropriate for the first stage of processing. Thus, rigid blind deconvolution in this first stage is unlikely to be optimal.

1.2 Pitch Perception

Processing voicing information in speech requires analyzing the harmonic structure associated with a quasi-periodic vocal driving function and might therefore be considered as an aspect of pitch perception.

In 1951, Licklider proposed a "duplex" theory [11] to account for many properties of pitch perception, including the perception of the missing fundamental (or residue pitch), as well as the pitch of modulated noise. Licklider envisioned neural machinery that measured the running temporal autocorrelation in each auditory frequency channel. The sensation of pitch, he proposed, is associated with the common periodicities observed across channels.

In 1984 Lyon was able to simulate an implementation of the duplex theory, labeling the graphic output a *correlogram* [12]. Since then, Meddis and colleagues [13] [14] have formalized the simulations and included a final stage that adds the running autocorrelations across each channel generating a *summary correlogram*. Cariani and Delgutte have also shown that similar processing of measured auditory-nerve impulses is sufficient to predict many classic pitch perception phenomena [2]. Other researchers have replaced the autocorrelation function with different mechanisms that measure the temporal intervals in each channel (e.g. [15] [4] [5]).

In general (and as shown in Licklider's original sketches achieved without the aid of computational simulation), simulations using these models provide a graphical output that correlates well with pitch. The time lag of the peak in the summary correlogram is usually found to be the reciprocal of the frequency of the perceived pitch and the height of the peak is often correlated with pitch salience. With few exceptions however, the models are not used to predict psychoacoustic just-noticeable-differences (jnds) with general stimuli. Together with the lack of a clearly identified physiological substrate for the implementation of the required timing measurements, this line of research remains somewhat of an "open-loop."

1.3 Perception of Amplitude Modulation

Processing voicing information in speech might also be thought of as a form of amplitude modulation perception.

In 1979, Viemeister applied a linear systems approach to the detection of acoustic envelope fluctuations [21]. His model was first fit to data describing the detection of sinusoidal amplitude modulation of wideband noise and then used to predict the detection of other harmonic envelopes. Motivated by the close relationship between standard deviation and autocorrelation, Viemeister's model used the standard deviation of a demodulated envelope as the statistic to predict human performance. Although this measure does not characterize the perceived pitch of the amplitude modulation, a more sophisticated simulation involving autocorrelation was not required to accurately fit the detection data. More recently, this model has been extended to predict other amplitude modulation detection data [19] [20].

In 1989, Houtgast measured modulation masking that suggested explicit neural modulation filtering [8]. Narrow-bandwidth noise modulators were found to mask the perception of sinusoidal modulators in a manner similar to the spectral masking of tones by narrow-band noises. Modulation tuning has also been measured physiologically [e.g., 9]. However, other modulation masking experiments, using sinusoids, have been less conclusive [19] [1]. Nonetheless, a model of modulation filtering has been implemented and shown to be correlated with many aspects of amplitude-modulation perception [3].

In essence, modulation filtering replaces the single low-pass filter in the envelope statistic model with a second bank of filters. The modulation filtering simulations also include a better approximation of auditory filtering than the single band-pass filter used in the envelope statistic model.

Therefore, there are at least three modeling approaches which may be helpful for analyzing the periodic envelope fluctuations in voiced speech: autocorrelation or interval-based temporal processing, the measurement of an envelope statistic and explicit modulation filtering. To choose among them, implementations of each were first fit to predict TMTF data and then each was used to predict the discrimination of voicing for strident fricatives in noise.

2. A Strident Fricative Case Study

Fricatives are generated by forcing air through a sufficiently narrow constriction in the vocal tract, resulting in a turbulent, noise-like source. With voiced fricatives the vocal folds also vibrate, adding low-frequency energy to the spectrum. The relative level of the first harmonic, compared with that of the adjacent vowel, has been shown to serve as an effective indicator of voicing distinctions for fricatives [18] [16].

2.1 Characterizing [s] and [z]

For our study, the strident fricatives [s] and [z], along with the vowels [a], [i] and [u] were recorded as CV syllables from four talkers. Figure 1 compares average log-magnitude spectral estimates for [s] and [z]. The voiced [z] has low-frequency energy not present in the [s].

Current ASR systems use the presence of low-frequency spectral energy to discriminate these sounds. However, there are situations where this particular spectral cue can be obscured (e.g. a high-pass channel or with a competing low-pass noise).

Figure 2 shows examples of the temporal waveform for [s] and [z], after each has been high-pass filtered above 3 kHz. Without low-frequency spectral components, the low-fre-

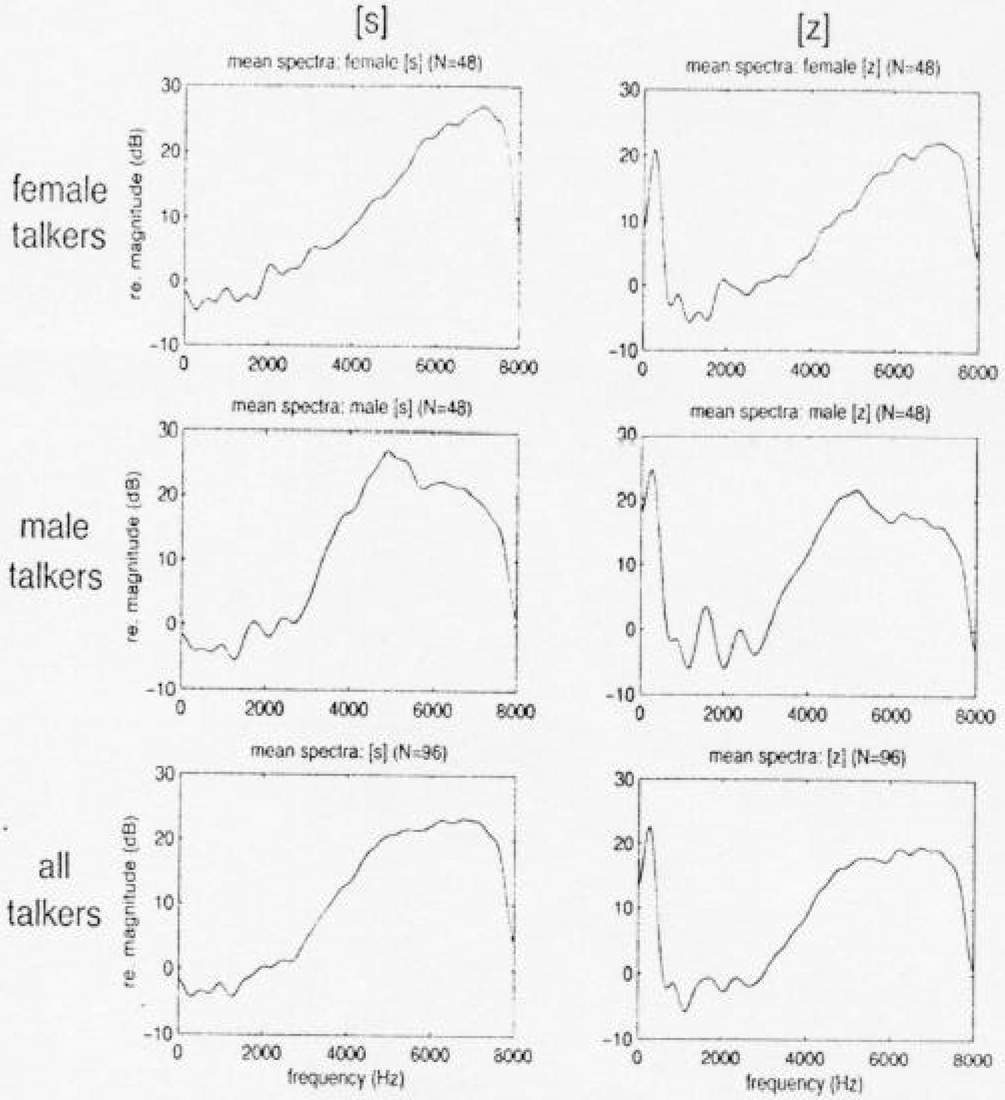

Figure 1 A comparison of average spectral estimates for [s] and [z] spoken by both male and female speakers.

quency pitch-rate information is represented in the envelope of the high-frequency, noise-like carrier. These figures provide evidence that the vibrating vocal folds can modulate the pressure source that drives the turbulence for a voiced fricative. The modulated noise source leads to a potentially redundant voicing cue in a spectral region with significant speech energy. ASR systems that integrate spectral estimates over multiple glottal periods do not distinguish such sounds, while human listeners can distinguish them even at low signal-to-noise ratios (see Section 4).

2.2 *Perceptual Measurements*

To measure the perceptual sensitivity to this potential voicing cue, the discrimination of these sounds was measured in wide-band noise. The syllable-initial fricatives were temporally isolated from the adjacent vowel, and high-pass filtered above 3 kHz. During the perceptual tests, tokens were centered within a one-second span of spectrally flat noise.

Adaptive tests [16] were used to track the perceptual discrimination of the isolated fricative as a function of SNR at two d' levels. For each trial, the subject was required to identify a randomly chosen token as either [s] or [z]. Feedback was provided. The initial SNR was sufficiently high that the fricatives were clearly distinguishable for all subjects. The SNR

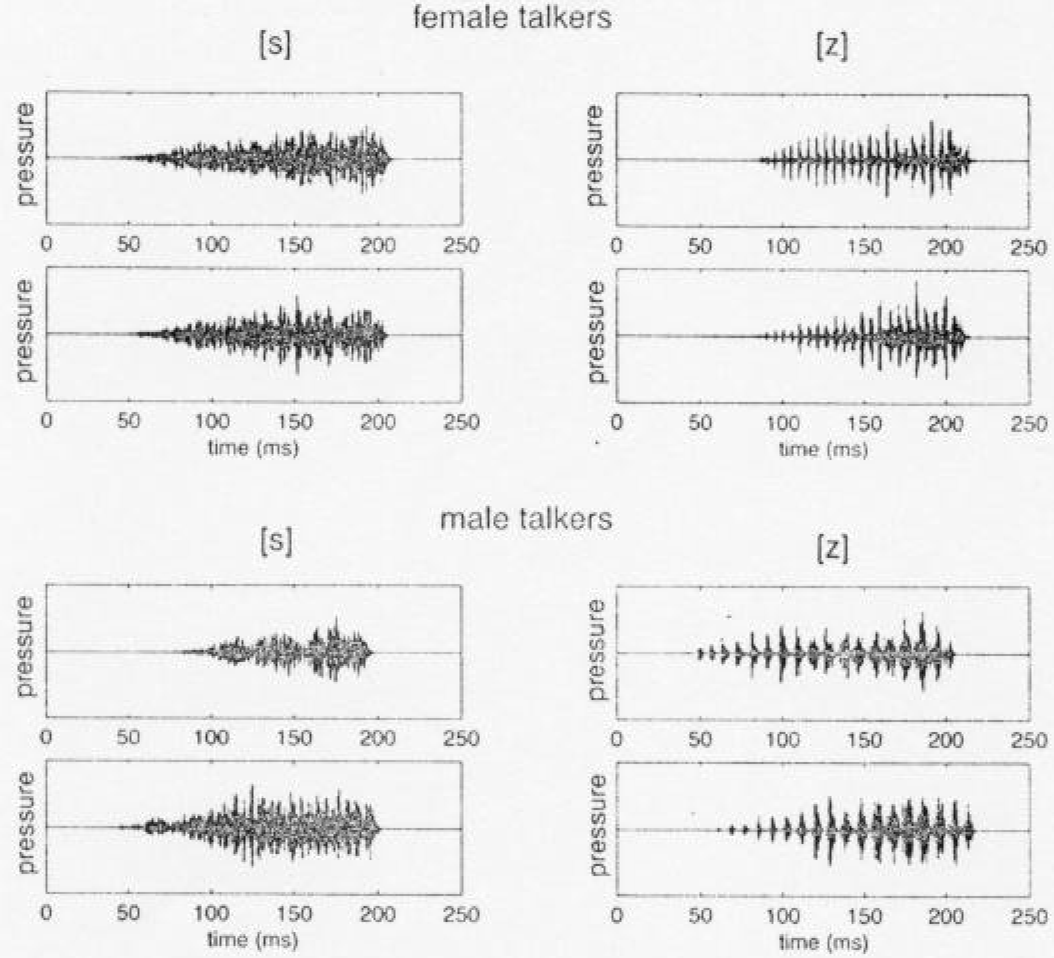

Figure 2 Examples of temporal waveforms after high-pass filtering.

was increased after an incorrect response and decreased after either two or three consecutively correct responses. (A reversal is defined as a change in the direction of the SNR step). The SNR step size started at 4 dB, and was reduced to 2 dB after the first reversal, and to 1 dB after the third. The average of the SNR at the next 6 reversals provided an initial threshold estimate. If the variance in this estimate was less than 2 dB, the measurements stopped, otherwise the experiment continued for up to 6 more reversals. The average of three such measurements provided a final threshold estimate for each subject. When 2 (or 3) correct responses are required, the threshold estimate converges to a 70.7% (or 79.4%) correct response rate. For this experiment, these correspond to d' values of 1.09, and 1.64, respectively. Four audiometrically normal subjects participated in the experiment. Average thresholds across these four subjects are shown together with model predictions in Figure 10 below.

3. AM-Detection Mechanisms

The task in this experiment requires detecting periodic envelope fluctuations, which become increasingly weak with the addition of noise. Perhaps the most direct approach is to model this perceptual process using an envelope statistic.

3.1 Envelope Statistic

Figure 3 shows a block diagram of the signal processing in an envelope-statistic model. This classical approach reduces auditory processing to the following steps: auditory filtering

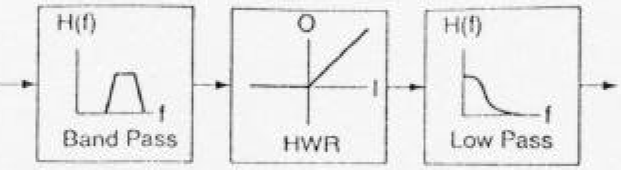

Figure 3 Schematic illustration of the envelope detection process used in the current study.

(measured along the basilar membrane), half-wave rectification (approximated in inner-hair-cell transduction) and low-pass filtering (computed throughout the higher levels of auditory processing). From an engineering perspective, the band-pass filter selects a channel, while the half-wave rectifier serves as a non-linearity that modulates the carrier down to DC, with the low-pass filter tracking the envelope.

The model's sensitivity to amplitude-modulated wideband noise increases with a broadening of the bandwidth in the initial filter, while the reduction of sensitivity with increasing envelope frequency is mostly determined by the final low-pass filter.

3.2 Modulation Filtering

A schematic overview of an implementation of modulation filtering is shown in Figure 4. Building from the envelope-detection processing above, the model includes multiple 4th-order gammatone filters [15] which provide a reasonable approximation of auditory filtering, and replaces the single low-pass filter with a second filterbank that analyzes the envelope spectrum.

The frequency response for the modulation filters used (Q_{3dB} of 2, and -12 dB DC gain) was adapted from [3]. For each filter the implementation used a second-order pole and a first-order (real) zero at DC. The distance of the zero to the unit circle was set to meet the DC specification. The resulting frequency responses are shown in Figure 5.

Both the modulation filtering and the envelope-detection model compute the magnitude of the fluctuations of the envelope of the acoustic waveform. As stated previously, the primary difference is that modulation filtering assumes a second, filtering stage tuned to different envelope modulation rates. Figure 6 compares the processing output of these two models to a noise carrier with no modulation, as well as one with 56% modulation [20 log(m) = 5 dB depth]. Although the standard deviation of the input is the same for the modulated and unmodulated cases, the outputs of both models exhibit relatively more fluctuation in the modulated case.

3.3 Correlational Analysis

An overview of the correlational analysis is shown in Figure 7. This is an implementation of Licklider's model [11] together with a final stage that adds correlation estimates across channels [13] [14]. The first stage is the same gammatone approximation of cochlear filtering, used above. The transduction stage includes half-wave rectification, low-pass filtering, and a 2nd-order Butterworth high-pass filter with a cut-off of 4 Hz. Running autocorrelations are computed in each filter channel and the results are summed across channels.

Our implementation of running autocorrelation for each channel involves two stages. First, the instantaneous product of the current input, and a version of the input delayed by the interval, τ, is computed for all time and all values of τ:

$$x_I(t,\tau) = x(t)\, x(t-\tau).$$

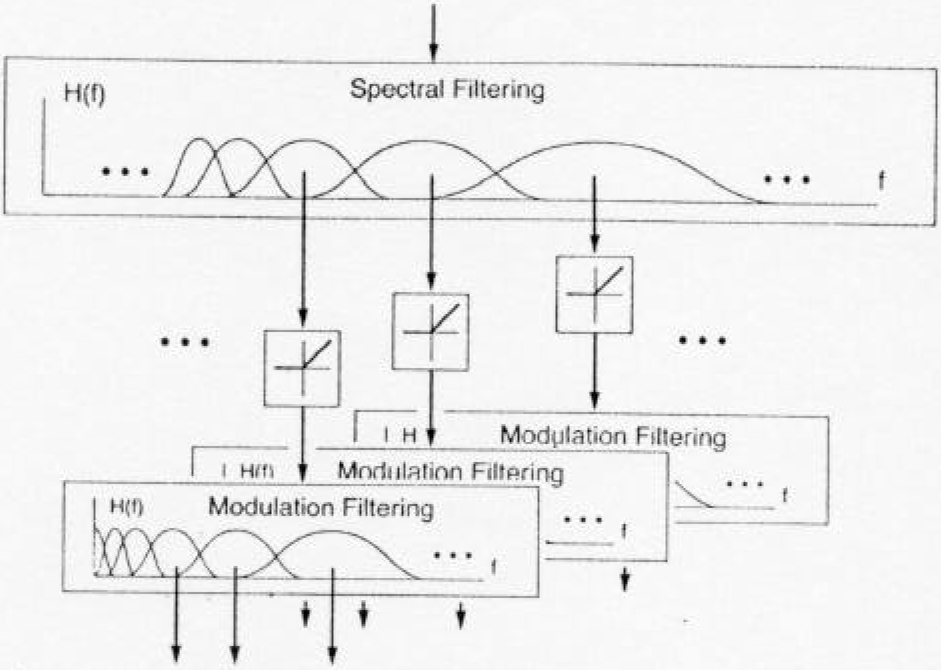

Figure 4 A schematic illustration of the modulation filtering performed in the current study.

Second, to form a running autocorrelation estimate, these sequences are low-pass filtered (for each value of τ) to below one half of the final correlation sampling rate:

$$x_2(t,\tau) = x_1(t,\tau) * h_{lpf}(t).$$

In the evaluations below, the correlation sampling rate was 25 Hz, and $h_{lpf}(t)$ was implemented as a 6th-order Butterworth filter with a -3 dB point at 10 Hz. That is, after the low-pass filter, the running autocorrelations were sampled every 40 ms and then summed across frequency channels to generate a sequence of summary correlogram estimates.

As described above, the position of the peak in the summary correlogram has often been shown to be correlated with the reciprocal of the perceived pitch (in units of frequency), although some models utilize the entire waveform of the summary correlogram [13] [14]. Our analysis represents a compromise between these two approaches. For each sample of the summary correlogram, our statistic is the maximum difference, across all delay values τ, between the summary correlogram values at delays of τ and $\tau/2$:

$$statistic = max\ [sc(\tau) - sc(\tau/2)],\ (0 < \tau < 20\ ms).$$

With a sinusoidal envelope, this difference peaks at a value of τ, equal to the period of the sinusoid. Figure 8 includes examples of this decision statistic using the same noise carrier, but with either no modulation or with 56% modulation (i. e., 5 dB depth) at 100 Hz. In the

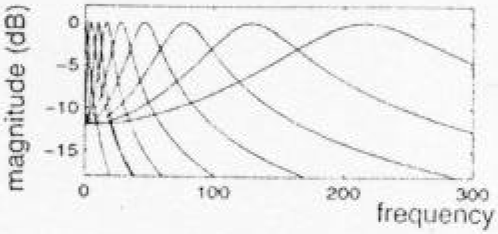

Figure 5 Responses of the modulation filterbank. Each filter is implemented using a complex pole and a real zero.

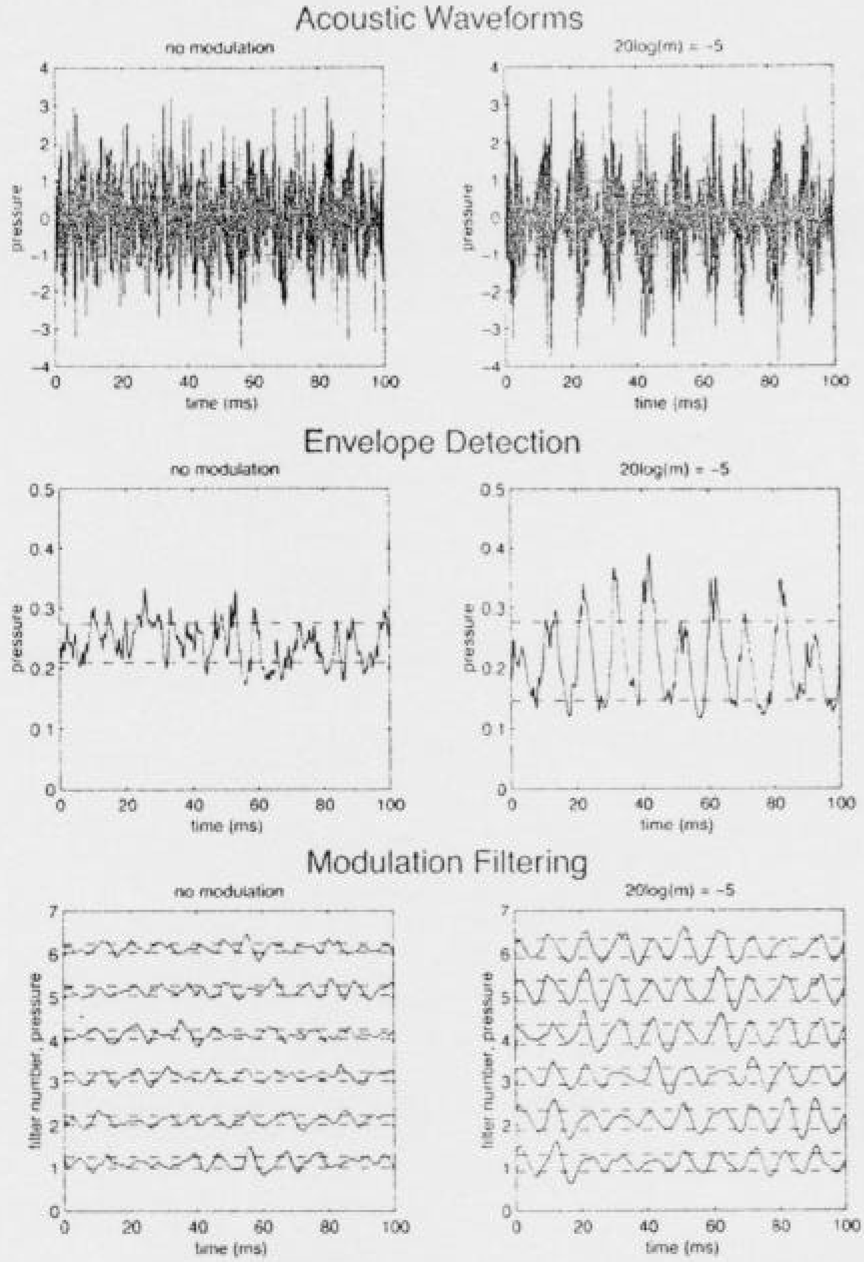

Figure 6 Comparisons of the amplitude modulation detection models. Dashed lines indicate standard deviations. The modulation filtering plots show the outputs of six auditory channels, each filtered by a modulation filter centered at 100 Hz.

modulated case, the first peak (after zero delay) in the summary correlogram occurs at the period of the modulation, 10 ms. When there is no modulation, the summary correlogram approximates an impulse. Adding the individual correlation estimated across channels reduces some of the variance; consistent modulation patterns across channels add together, while inconsistent ones generally cancel each other. However, considerable variation remains across summary correlogram samples (shown in the lower half of Figure 8) due to the stochastic nature of the carrier.

4. Comparing Predictions

The temporal modulation transfer function (TMTF) is a measure of auditory sensitivity to amplitude modulation as a function of modulation frequency. More specifically, the mini-

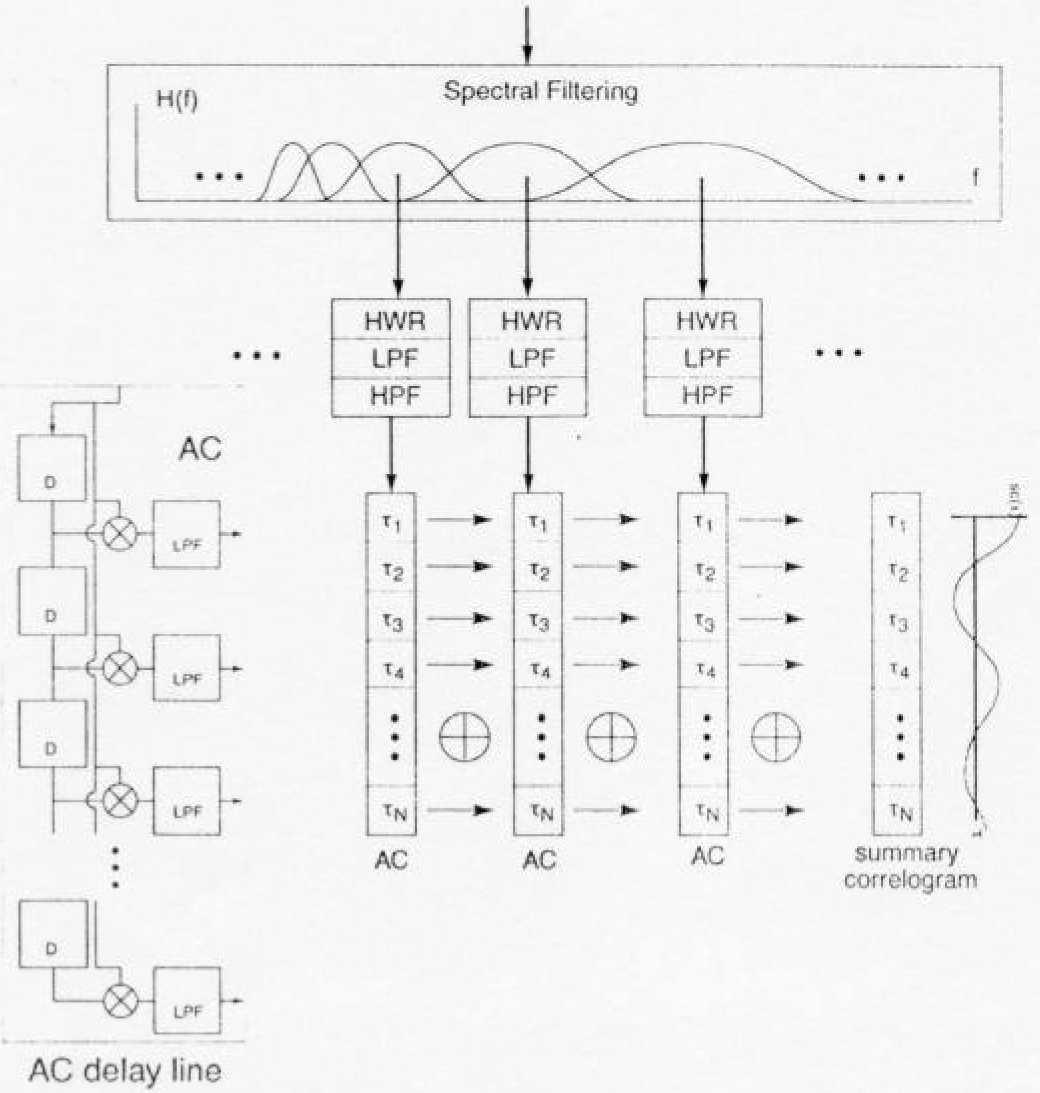

Figure 7 Overview of the correlational processing. Inset shows autocorrelation delay-line detail.

mum detectable sinusoidal amplitude modulation depth is typically measured as a function of modulation frequency using wide-band noise carriers.

Each of the three models was initially adjusted to predict TMTF measurements derived from previous studies [20] [3]. The resulting models were then used to predict the discrimination thresholds for the high-pass filtered [s] and [z] tokens in noise. Because the natural fricatives are non-stationary all three models were evaluated using multiple measurements in time (or multiple "looks") [22].

For the envelope-statistic model, the best match was found using an initial filter bandwidth of 3 kHz, centered at 5.5 kHz. With these parameters, the filter approximated a matched-filter for the high-pass filtered [s] and [z] segments. The low-pass filter was a 1st-order Butterworth with a cut-off of 90 Hz. The normalized fourth-moment statistic [19] [20] was used.

To obtain multiple measurements in time, the output of the envelope detection mechanism was segmented using partially overlapping, 50-ms windows that had 10-ms raised-cosine onset and offsets, as well as a 30-ms steady-state center. The windows were incremented by 40 ms. The window length was chosen to ensure multiple periods in each window for the pitch-frequency range of interest. By modulating the DC offset in the envelope, the

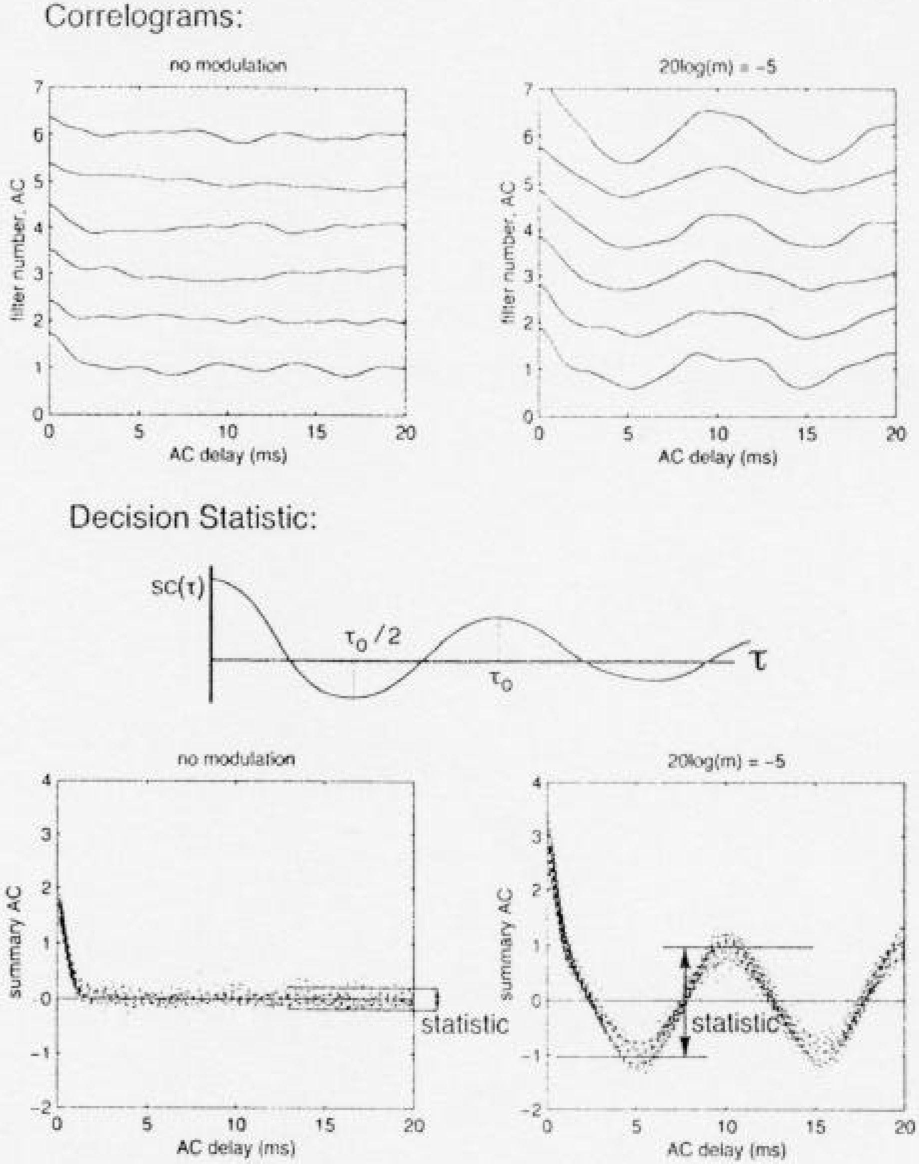

Figure 8 Samples of the correlogram output and super-imposed examples of the summary correlogram decision statistic. Input signals are the same as in Figure 6.

shape of the window can dominate measurements using the standard deviation or the fourth-moment. Therefore, the DC offset for each 50-ms window was removed before weighting by the raised-cosine and then added back before computing the statistic.

Threshold predictions were obtained by using the difference in the decision statistic in signal and non-signal intervals over 100 simulations in order to estimate d' for each "look." Assuming independence of individual measurements, a total detection d' was estimated as the length of a d' vector containing all looks [7]. With a stimulus duration of 500 ms used for the TMTF data, the vector included 12 elements (or looks). A line was fit to the log of total d' estimates as a function of the log of the modulation depth. From this line, the modulation threshold was estimated from the point where the line crossed the d' threshold of 1.26 tracked in the perceptual TMTF measurements [20] [3].

With the modulation filtering and correlation models, the initial filtering stage was six, 4th-order gammatone filters with center frequencies range between 4.28 Hz and 6.97 kHz. Filters overlapped at their half-power points, and the bandwidths were set using the equation described in [6]. To predict the TMTF data using modulation filtering only the modulation

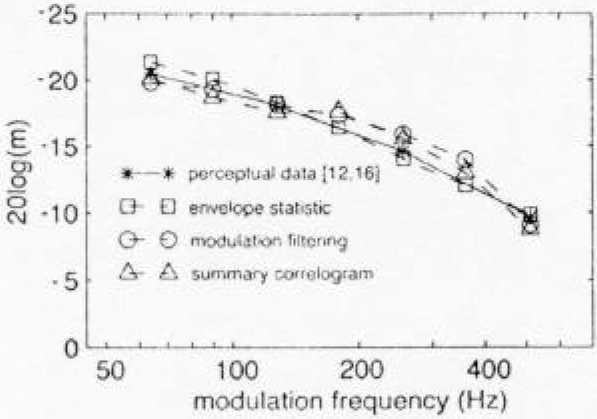

Figure 9 Three predictions of TMTF data: *m* is the modulation depth; perceptual data is an average from [20] and [3].

filter tuned to the probe envelope frequency was considered. When predicting the fricative data two modulation filters centered at 120 Hz and 200 Hz were used. The windowing applied to the envelope-detection simulations was also used for the modulation filtering. The standard deviation was the measured statistic.

As observed previously [3], the modulation filtering was too sensitive to predict human performance without adding a large amount of internal noise. To obtain the best match to the TMTF data, internal noise was added both before and after modulation filtering.

Using the correlation model, the peak distance statistic described above was measured every 40 ms for the summary correlogram. To approximate the shape of the TMTF data, the first-order, low-pass filter was used with a cut-off frequency at 280 Hz.

TMTF threshold predictions for all three models are shown in Figure 9. Each model provides a reasonable prediction across this frequency range. Predicting the voicing detection thresholds for the natural, non-stationary, fricatives in noise required finding the fricatives (or more specifically finding the voicing in the fricative) within the 1-second interval of noise. For all model predictions below, only the three consecutive temporal segments that maximized the difference from the background noise were analyzed, providing three temporal looks per token. Total *d'* values were then estimated as a function of SNR.

Figure 10 shows the *d'* estimates for each model's prediction of the discrimination of the high-pass filtered [s] and [z] tokens in noise. The model based on correlations provided the best prediction.

5. Modeling Implications

The envelope statistic was not sufficient, by itself, to discriminate reliably between [s] and [z] (even at relatively high SNR values) because this measurement does not distinguish the periodic voicing cues in [z] from the aperiodic fluctuations in [s]. Both the modulation filtering and the autocorrelation processing include specific modulation tuning and as a result more accurately fit the observed data.

Reasons for the difference in performance between these two models are less clear, and could be specific to these simulations. By reducing the amount of internal noise, the modulation filtering model provides a better estimate of the [s] and [z] data, but over-estimates the TMTF sensitivity. One primary difference is that the autocorrelation mechanism integrates correlation estimates across frequency, while the modulation-filtering simulations use the more general assumption that each output corresponds to an independent measurement. Inte-

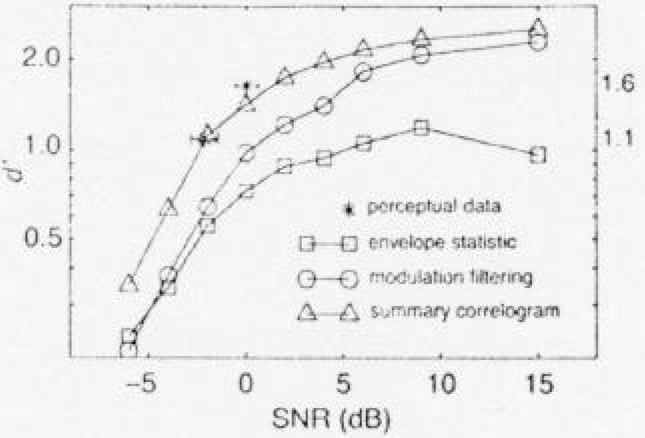

Figure 10 Discriminating high-pass filtered [s] and [z]: data are an average across four subjects.

grating correlation estimates across frequency channels de-emphasizes envelope components uncorrelated across frequency in favor of correlated components. Another difference is that the correlation simulations uses a low-pass filter to limit sensitivity, while the modulation simulation incorporates internal noise.

It is interesting to note that if the auditory system does include a cross-channel interval-based representation, redundancies in this representation are likely to make it inefficient to maintain across many regions of the pathway. Efficient decorrelation of the (potentially smooth and periodic) summary correlogram might approximate a cosine transform. Such periodic transformations exist in other perceptual systems [23]. In this case the decorrelated representation would have many of the properties of the (demodulated) output of a modulation filterbank. The difference is that the envelope analyzed is first processed to identify common correlations across a broad frequency range.

6. Conclusion

This chapter has identified a secondary temporal cue that can reliably distinguish between [s] and [z] on the basis of voicing. This amplitude-modulation cue had not been identified in previous studies of voiced fricatives [18] [16]. Once the cue has been identified it is not clear what processing should be used to reliably extract it. Three possibilities were investigated in this study.

While cross-channel, interval-based processing has been quite successful in predicting many properties of pitch perception, we have shown that these mechanisms can also predict TMTF thresholds and the detection of voicing for high-pass filtered fricatives in noise. Simulations using envelope-statistic and modulation-filtering models fit TMTF data, but do not predict the isolated speech data.

Acknowledgements

We thank the subjects for their efforts, and Jim Hant for helpful discussions and for his assistance in running some of the experiments. This research was supported by NIDCD grant #DC02033 and the NSF.

References

[1] Bacon, S. P. and Grantham, D. W. "Modulation masking: Effects of modulation frequency, depth, and phase." *J. Acoust. Soc. Am.*, 85: 2575–2580, 1989.

[2] Cariani, P. A. and Delgutte, B. "Neural correlates of the pitch of complex tones: I. Pitch and pitch salience." *J. Neurophysiol.*, 76: 1698–1716, 1996.

[3] Dau, T., Kollmeier, B. and Kohlrausch, A. "Modeling auditory processing of amplitude modulation. I–II." *J. Acoust. Soc. Am.*, 102: 2892–2919, 1997.

[4] de Cheveigne, A. "Cancellation model of pitch perception." *J. Acoust. Soc. Am.*, 103: 1261–1271, 1998.

[5] Ghitza, O. "Auditory nerve representations as a basis for speech processing." In *Advances in Speech Processing*, S. Furui, M. Sondhi (eds.), New York: Marcel Dekker, pp. 453–485, 1991.

[6] Glasberg, B. R. and Moore, B. C. J. "Derivation of auditory filter shapes from notched-noise data." *Hearing Res.*, 47: 103–138, 1990.

[7] Green, D. M. and Swets, J. A. *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

[8] Houtgast, T. "Frequency selectivity in amplitude-modulation detection." *J. Acoust. Soc. Am.*, 85: 1676–1680, 1989.

[9] Langner, G. "Periodicity coding in the auditory system." *Hearing Res.* 60: 115–142, 1992.

[10] Levitt, H. "Transformed up-down methods in psychoacoustics." *J. Acoust. Soc. Am.* 49: 467–477, 1971.

[11] Licklider, J. C. R. "A duplex theory of pitch perception." *Experientia*, 7: 128–134, 1951.

[12] Lyon, R. F. "Computational models of neural auditory processing." *Proc. IEEE ICASSP*, 36.1: 1–4, 1984.

[13] Meddis, R. and Hewitt, M. J. "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification." *J. Acoust. Soc. Am.*, 89: 2866–2882, 1991.

[14] Meddis, R. and O'Mard, L. "A unitary model of pitch perception." *J. Acoust. Soc. Am.*, 102: 1811–1820, 1997.

[15] Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand, M."Complex Sounds and Auditory Images." In *Auditory Physiology and Perception*, Y. Cazals and K. Horner (eds.), Oxford: Pergamon Press, pp. 429–446, 1992.

[16] Pirello, K., Blumstein, S. and Kurowski, K. "The characteristics of voicing in syllable-initial fricatives in American English." *J. Acoust. Soc. Am.*, 101: 3754–3765, 1997.

[17] Rabiner, L., Juang, B. H. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[18] Stevens, K. N., Blumstein, S., Glicksman, L.,Burton, M., Kurowski, K."Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters." *J. Acoust. Soc. Am.*, 91: 2979–3000, 1992.

[19] Strickland, E. and Viemeister, N. "Cues for discrimination of envelopes." *J. Acoust. Soc. Am.*, 99: 3638–3646, 1996.

[20] Strickland, E. and Viemeister, N."The effects of frequency region and bandwidth on the temporal modulation transfer function." *J. Acoust. Soc. Am.*, 102: 1799–1810, 1997.

[21] Viemeister, N. "Temporal modulation transfer function based on modulation thresholds." *J. Acoust. Soc. Am.*, 66: 1364–1380, 1979.

[22] Viemeister, N., Wakefield, G. "Temporal integration and multiple looks." *J. Acoust. Soc. Am.* 90: 858–865, 1991.

[23] Wang, K. and Shamma, S. "Self-normalization and noise-robustness in early auditory representations." *IEEE Trans. Speech. Aud. Proc.*, 2.3: 412–435, 1994.