Published in final edited form as:
Ear Hear. 2008 October ; 29(5): 800–813. doi:10.1097/AUD.0b013e31817e73ef.

Exploring the role of the modulation spectrum in phoneme recognition

Frederick Gallun, Ph.D.
VA RR&D National Center for Rehabilitative Auditory Research Portland VA Medical Center Portland, OR

Pamela Souza, Ph.D., CCC-A
Department of Speech and Hearing Sciences University of Washington Seattle, WA

Abstract

Objectives—The ability of human listeners to identify consonants (presented as nonsense syllables) on the basis of primarily temporal information was compared with the predictions of a simple model based on the amplitude modulation spectra of the stimuli calculated for six octave-spaced carrier frequencies (250-8000 Hz) and six octave-spaced amplitude modulation frequencies (1-32 Hz).

Design—The listeners and the model were presented with sixteen phonemes each spoken by four different talkers processed so that one, two, four or eight bands of spectral information remained. The average modulation spectrum of each of the processed phonemes was extracted and similarity across phonemes was calculated by the use of a spectral correlation index (SCI).

Results—The similarity of the modulation spectra across phonemes as assessed by the SCI was a strong predictor of the confusions made by human listeners.

Conclusions—This result suggests that a sparse set of time-averaged patterns of modulation energy can capture a meaningful aspect of the information listeners use to distinguish among speech signals.

Introduction

Clinicians are accustomed to thinking about sound in terms of its level as a function of frequency, where the frequencies of interest vary from about 250 Hz to about 6000 Hz, corresponding to the expected range of speech information. The frequency spectrum can be measured over a long duration, as for the long-term average speech spectrum (Cox & Moore, 1988); over a short duration, as for an individual speech sound (Pittman, Stelmachowicz, Lewis, & Hoover, 2003); or over sequential short-duration periods, as for a spectrogram (see Figure 1). Such a view is the basis for many clinical tools, including hearing aid prescriptive methods intended to provide audibility across frequency (Scollie et al., 2005), “count the dot” audiograms (Mueller & Killion, 1990), and the Speech Intelligibility Index (ANSI, 1997).

It has been proposed that the information carried by the speech signal can also be represented by the changes in envelope that occur over time, or the amplitude-modulation spectrum (referred to here as simply the modulation spectrum). In this concept, speech information is conveyed by a composite of modulations at multiple rates, superimposed on a carrier signal. Thus, a series of short-duration snapshots of the signal can be said to contain both the frequency information (what we call “carrier frequency” information) and amplitude-modulation information (what we call “modulation frequency” information). An analysis that focuses on only the long-term frequency spectrum de-emphasizes changes in amplitude envelope, while an analysis focusing on the long-term modulation spectrum emphasizes the envelope changes while de-emphasizing the carrier frequency content. Historically, the long-term frequency content has been emphasized in terms of speech intelligibility, but recent developments in the

field of cochlear implants have begun to suggest that even when long-term frequency content is held fixed, listeners can extract substantial information from the modulation information alone (e.g., Shannon, Zeng, Kamath, Wygonski, and Ekelid, 1995).

In addition to the many practical demonstrations provided by the growing numbers of successful cochlear implant users, there is also convincing evidence from the psychophysical literature that the auditory system is tuned not only in the carrier frequency domain, but in the modulation frequency domain as well (Houtgast, 1989; Bacon and Grantham, 1989; Yost and Sheft, 1989). These data have led to the development of new models (Dau, Kollmeier and Kohlrausch, 1997a, 1997b; Ewert and Dau, 2001; Chi, Gao, Guyten, Ru, and Shamma, 1999; Chi, Ru and Shamma, 2005) that represent the auditory system as composed of two cascaded sets of band-pass filters, the first in carrier frequency and the second in modulation frequency. Essentially, these models argue that the carrier frequency energy that is extracted by a long-term spectral analysis (such as would be used to compute the average speech spectrum) should be viewed as a dynamic signal that is changing over time. Rather than averaging these changes in level over time, the output of each carrier-frequency filter (the “critical bands” of Fletcher, 1940) is fed to a bank of amplitude-modulation filters and it is the output of this second cascade of filters that provides the information necessary to understand speech. Notice that this does not discard the carrier-frequency information, but rather transforms it to extract the dynamic modulations within each band.

Modulation frequency bands have been measured psychophysically for modulated tones and modulated noise and the rates of modulation to which humans are most sensitive vary from roughly .5 Hz to about 64 Hz, with a peak of sensitivity at 4 Hz (Chi et al., 1999; Arai et al., 1999). This already suggests a relationship between modulation sensitivity and the extraction of speech information because this peak sensitivity corresponds to a duration of 250 ms, which is quite close to a common syllable rate for speech (Arai and Greenberg, 1997). While it may seem surprising to assert that frequencies well below the limits of acoustic sensitivity are a principle information-bearing component of speech, this idea has actually been around for many years. Dudley (1939) first described his “vocoder” as an instrument that functioned by using a code that involved “modulation processes of the true message-bearing waves, which, however, by themselves, are inaudible” (p.177). It is in this spirit that we refer to the audible frequencies in speech as the “carrier frequencies” and devote most of our effort to determining how best to represent the “true message-bearing waves”, which are the modulations imposed upon the carrier frequencies. This is not, of course, to say that the carrier frequencies are unimportant. Obviously, energy in the audible frequency range is crucial for conveying the modulations. At the limit, however, a speech signal containing no modulation would simply be a set of harmonically-related tones.

Modulation spectra have been successfully applied to audio coding (Vinton & Atlas, 2001) and automatic speech recognition (Greenberg & Kingsbury, 1997; Hermansky, 1997; Kanedera, Arai, Hermansky and Pavel, 1999), but most of these applications have required the time-varying amplitude-modulation waveform rather than a time-averaged modulation spectrum. Indeed, for *reconstruction* of the speech signal, Atlas (Atlas, Li, & Thompson, 2004) has cautioned that calculation of the complex modulation spectrum consists of both an amplitude component and a phase component, and that discarding the phase component can result in audible artifacts. Greenberg and colleagues (Greenberg, Arai, & Silipo, 1998; Arai & Greenberg, 1998; Greenberg & Arai, 2001; Silipo, Greenberg, & Arai, 1999) produced results that support this conclusion, as they demonstrated that shifting the modulation spectra relative to one another in time (which amounts to changing the phase) degrades recognition. Despite these findings, the analysis used in the current work (described below) does not include modulation phase and thus will fail to capture the role of the relative (and time-varying) phases of modulation frequencies in phoneme recognition by human listeners. This simplification was

chosen in order to focus on the question of what information is conveyed by the modulation spectrum and the extent to which changes in the time-averaged modulation spectrum can serve as a concise way of describing changes in modulation information introduced by a device such as a hearing aid or a cochlear implant or by differences in speaker or speaking style. Some previous workers in this area (e.g., Green, Katiri, Faulkner, & Rosen, 2007; Krause & Braida, 2004) have used the modulation spectrum as a measure of changes between different speech conditions, but the amount or type of information contained in the modulation spectrum was not assessed. In this view, the modulation spectrum is considered as a unique descriptor of a particular speech segment. The focus of this paper is to assess the ability of the modulation amplitude spectrum alone (without a measure of modulation phase) to accurately predict human performance in a case where primarily modulation information is retained in the stimulus.

In order to examine the role of the modulation spectrum in speech recognition using a simple and easily implemented analysis method, a phoneme-specific analysis was completed. To support this analysis, processed speech stimuli were generated and identification scores were obtained for a group of young listeners with normal hearing. These processed speech stimuli (described below) varied in the extent to which independent modulation information was retained across carrier frequencies and the processing equated the long-term frequency spectra across all of the speech tokens. A simple, time-averaged measure of modulation spectrum information was obtained for each speech token and correlations across tokens were used to predict the similarity, and thus the likelihood of confusions, between the tokens. The degree to which the correlations in modulation spectra predict error patterns for the human listeners was used to assess the amount of speech information that was captured by the long-term amplitude-modulation spectrum.

Method

Participants

Ten participants aged 22-30 years (mean age 25.8 years) were recruited for the study. All participants had normal hearing, defined as pure-tone thresholds of 20 dB HL or better (re: ANSI, 2004) at octave frequencies between .25 and 8 kHz bilaterally, and spoke English as their first or primary language. One ear of each participant was randomly selected for testing. All participants were paid for their participation and all procedures were reviewed and approved by the University of Washington Institutional Review Board.

Stimuli

Test stimuli were a set of 16 vowel-consonant-vowel syllables, containing one of the following consonants /b, d, g, p, t, k, f, θ, s, ʃ, v, ð, z, Z, m, n/ in an /aCa/ context. Each token was produced by four talkers (two male and two female) without a carrier phrase for a total of 64 test items. All tokens were recorded at a 44.1 kHz sampling rate with 16 bit resolution. Syllables were normalized such that the vowel level of each was approximately equal.

Four processed-speech test conditions were created: one-channel, two-channel, four-channel, eight-channel. An unprocessed condition was included as a control comparison. To create the processed conditions, each unprocessed speech token was digitally filtered into channels. The lower cutoff frequency for the lowest frequency channel was 176 Hz and the upper cutoff frequency for the highest frequency channel was 7168 Hz. Intermediate cutoff frequencies were based on logarithmically 1 spaced 1000 point FIR bandpass filters (Table 1). The filtered segments were processed to limit spectral information by randomly multiplying each digital sample by a +1 or -1 (Schroeder, 1968). Unlike other methods of creating vocoded signals which employ envelope extraction, no smoothing filter was used to restrict the range of envelope frequencies. Envelope frequencies were therefore available up to the limits of the

listener's ability to detect such cues. Next, each segment was refiltered using the original filter settings and amplified with gain appropriate to correct for the power loss of the second filtering. The filtered segments were digitally mixed. As described below, the resulting syllables retained modulation spectrum cues but obscured the carrier frequency information to varying degrees. The one-channel signal provided no carrier frequency information (i.e., the carrier was a broad-band noise). The two-, four- and eight-channel signals provided varying amount of carrier frequency information, albeit substantially less than available in the unprocessed signals.

Measurement of modulation spectra

The modulation spectrum of each signal was calculated and represented as the energy at the output of a bank of six octave-band modulation filters (centered at 1, 2, 4, 8, 16 and 32 Hz) for each of six octave-band carrier frequency filters (centered at 250, 500, 1000, 2000, 4000 and 8000 Hz). These thirty-six values (six modulation frequencies for each of the six carrier frequencies) are a limited representation of the modulation spectra for each of the stimuli, and there is no information about the relative phases of the modulation across channels. The calculation of modulation energy across bands was similar to the method described in Krause and Braida (2004). The first step limited the signals in the carrier frequency domain by filtering into the six octave-wide bands. No attempt was made to use filtering that matched human auditory filters in order to test the simplest version of the model and only include further complexity as required by the data. The second step in the extraction of modulation information limited the signals in the modulation frequency domain by half-wave rectifying each of the band-filtered signals and then passing the resulting signal through a 50 Hz low-pass filter. The removal of modulation information above 50 Hz is another simplification used in the measurement operation that does not reflect the limits of human processing (Chi et al., 1999) and that could be changed in future versions of the model. Alternatively, however, such filtering is what would be suggested by the results of Arai et al. (1999) as well as Drullman et al. (1994a;b). The first two processing steps (filtering in the carrier frequency domain and then half-wave rectifying and low-pass filtering) transformed a waveform that could be presented over headphones into a set of six envelope signals with no frequency content above 50 Hz.

The third processing stage resulted in an estimate of the frequency content of these envelopes (i.e., the “modulation spectra”). Each envelope signal was downsampled to a sampling rate of 1000 Hz and the resulting signal was submitted to a Fast-Fourier Transform (FFT). Prior to the FFT, the signal was zero-padded such that the duration of the signal was extended to five seconds, thus allowing a frequency resolution in the FFT of .2 Hz. In order to analyze the output of the FFT, the energy in each .2 Hz bin between .2 and 64 Hz was summed with the energy in adjacent bins. The choice of which bins to sum was made such that the summed energy was obtained for the equivalent of six rectangular filters with bandwidths of one octave and center frequencies stretching from 1 Hz to 32 Hz. This summed energy value was then divided by the energy in the 0 Hz or DC bin in order to provide a normalized “modulation index” value, which indicates the relative amount of modulation in each filter. This normalization ensures that a sinusoidal modulation at a given rate and a modulation depth of 100% would yield a modulation index of 1 for the filter containing that modulation rate. It was at this final stage that the relative phases of the modulation patterns and the fine-grained differences in modulation patterns (between 30 Hz and 31 Hz, for example) were discarded.

Three examples in which the modulation energy is contained within a single filter are shown in Fig. 2. On the left are three waveforms that are 100% sinusoidally amplitude modulated at rates of 1 Hz, 4 Hz and 16 Hz. On the right are the outputs of the six modulation filters and it can be seen that for each the modulation energy is restricted only to the appropriate filter.

As each of the example signals represented in Fig. 2 had a single carrier frequency (1000 Hz), there is no need to represent the modulation in the other five carrier frequency bands. For the

signals used in the behavioral experiment, however, the differences in modulation across carrier frequencies are likely to play an important role in identification. Figure 3 shows how the unprocessed phoneme shown in Fig. 1 was filtered into six carrier bands, three of which are shown in the panels on the left. The resulting distribution of modulation energy across the filters for each carrier band is shown in the panels on the right. Although the maximum value for any one frequency component is restricted to 1 due to the normalization process, the fact that the analysis sums the energy in octave-band ranges around the center frequencies of the modulation filters means that the total energy can exceed values of 1.

Procedures

For behavioral testing, each subject was seated in a double-walled sound booth. The syllables were presented monaurally (single channel fed to one earphone) through an insert earphone (Etymotic ER2) at 65 dB SPL. The 16 syllable choices were displayed on a touch screen in front of the subject. After each syllable was presented, the subject selected the syllable heard. To become familiar with the task, each subject began with a practice set containing 64 unprocessed tokens (16 consonants × 4 talkers). After practice, the subject completed two sets (64 tokens each) for each of the five conditions (one-, two-, four-, eight-channels and unprocessed). Presentation order of the five conditions was randomized. Within each set of 64 tokens, presentation order of the tokens was randomly selected without replacement. Results consisted of an overall percent correct score and a confusion matrix for each condition.

Results

Modulation spectra

The main question that motivated the modulation spectrum analysis was whether or not the sixteen phonemes could each be considered to have a “signature” modulation spectrum. If so, it was hypothesized that the similarity of this signature across phonemes would be a good predictor of the confusions that human listeners would experience when presented with these stimuli. In order to address both of these issues, a new measure of modulation similarity was developed. This new measure, which will be referred to as the spectral correlation index (SCI) is based on the correlation of the thirty-six values (six modulation frequencies by six carrier frequencies) across various stimuli. For two signals, the SCI is defined to be the Pearson correlation value (r) between the modulation values associated with each signal. Initial analyses showed that the SCI across phonemes was not consistent across talkers, due primarily to variability in rate of production. While the issue of talker variability is clearly important (and is to be examined in detail in future work), in this case it was a nuisance variable. Since the listeners were required to perform the task on the basis of all four examples of each phoneme, it was decided that the SCI analysis should be conducted across all four as well. This was accomplished by concatenating the thirty-six values from all four examples into a single vector of one-hundred forty-four points and correlating these new vectors for every pair of phonemes. Thus, a single correlation was obtained for each pairing of phonemes. The name SCI was retained for referring to this analysis, which led to overall higher SCI values across phonemes, but still provided a wide enough range of SCI values that comparisons could be made with the behavioral data.

In order to address the question of whether or not each phoneme has a unique “signature” modulation spectrum, the SCI values were compared for the five levels of processing that were applied to the phonemes (one-channel, two-channel, four-channel, eight-channel, and unprocessed). Each SCI value was thus the correlation of two 2,304 point vectors (36 modulation values for each of 16 phonemes for each of 4 talkers). Because the modulation spectrum analysis was limited to six bands, it was predicted that the eight-channel and the unprocessed signals would be essentially identical. SCI values calculated for spectra

concatenated across all phonemes and all talkers for a given level of processing (see Table 2) showed that this was an accurate prediction, with a value of .96. Interestingly, however, the SCI between the unprocessed stimuli and the one-channel stimuli was still .87. This result could be interpreted to mean that in terms of the modulation spectra (as calculated here) the majority of the temporal information for these aCa stimuli is retained in even a one-channel simulation. An alternative interpretation, however, is that even though the correlation is quite high, human listeners are sensitive to small differences in modulation spectrum and so would find the differences that lead to an SCI of only .87 significant enough that ability to identify the phoneme would be reduced.

This second interpretation is supported by the range of SCI values that were obtained for the sixteen different phonemes combined across talkers but divided by type of processing. As there were similar ranges and distributions of SCI values for all five types of processing, the values for the unprocessed stimuli will be used as examples. For 107 of 120 unique comparisons, SCI values fell in the range between .58 and .90. The five values at or above .90 were for aθa and afa (.96), aka and ata (.95), ada and aba (.92), apa and ata (.90) and aθa and asa (.90). The eight values at or below .58 were for ada and aza (.58), asa and ada (.58), apa and aza (.57), ada and aʃa (.57), aða and apa (.56), asa and aba (.56), aza and aga (.53), and asa and aga (.52). The complete set of SCI values for the unprocessed stimuli appears in Table 3.

Behavioral data

Because none of the participants had prior test experience with these signals, data were first analyzed for potential learning effects. Overall proportion correct for the first and second sets are shown in Figure 4. There was no significant difference in scores between the first and second set for any test condition, $F(1,45)=.74$, $p=.393$. Accordingly, results from both sets were collapsed for analysis. Thus, final data for each subject is based on 128 tokens (16 consonants × 4 talkers × 2 sets) per condition.

For analysis of error patterns, results from all subjects were compiled into a master confusion matrix for each test condition (Appendices A-E). We were most interested in whether phonemes having similar modulation spectra were more likely to be confused with one another, particularly in the absence of frequency carrier information. To that end, the confusions for each presented phoneme were translated into error probabilities by dividing the number of times each phoneme was chosen by the total number of choices (80). Phonemes that were more similar to each other in modulation spectra were more easily confused. Figures 5 and 6 illustrate this effect for the one-channel condition, which is the condition with the least carrier frequency information (and thus the one where listeners must depend the most on the modulation properties of the sound). Results are shown for tokens /afa/ in Figure 5 and /ama/ in Figure 6. In each figure, the top panel represents the degree of similarity in modulation spectra and the bottom panel represents the probability of making a specific error. For example, the modulation properties of /afa/ are most like those of /aʃa/, /asa/, and /aθa/, as represented by the high values in the top panel of Figure 5. The 1.0 value for /f/ simply reflects the correlation of that phoneme with itself. As seen in the bottom panel of Figure 5, the most common errors were /aba/, /asa/, and /aθa/. A similar effect is seen for /ama/ in Figure 6. Note also that these are not the “traditional” or expected errors in speech recognition; for example, we might expect that /ama/ would be confused with /ana/ more than /aʃa/, but it was not.

To assess this relationship mathematically, bivariate correlations were completed between the modulation similarity and the probability of errors in each test condition, using the master confusion matrices created by collapsing across listener responses. Only errors that occurred more than 5% of the time were included in the analysis. Results are plotted in Figure 7 and summarized in Table 4. In all conditions, the relationship between the modulation spectrum and the probability of error was significant. Listeners were more likely to confuse those

phonemes which had similar modulation spectra, and unlikely to choose phonemes with dissimilar modulation spectra. The data in Figure 7 also reflect a number of instances where two phonemes have similar modulation spectrum but are not confused. This probably reflects that the two phonemes could be distinguished using information available to listeners beyond the modulation properties of the sound; or, at least, information not reflected in this particular quantification of the modulation spectrum. For example, if two phonemes are similar in modulation spectra but differ in the relative timing of particular modulations (such as rising frequency glide versus a falling frequency glide), this method of capturing the modulation information would predict confusions that listeners might not experience. Figure 7 also shows that the overall probability of making an identification error decreased with increasing numbers of channels, but the correlation between the SCI and the listener errors increased (see Table 4).

We expected that conversion of the signals to vocoding would eliminate fine-structure cues but retain cues that could be quantified by the modulation spectrum. According to Rosen (1992), modulation rates between 2 and 50 Hz should allow the listener to distinguish consonant manner (e.g., the stop consonant /t/ versus the fricative consonant /s/) and voicing (e.g., /s/ vs. /z/). In theory, loss of fine-structure cues should result in lower transmission of consonant place such that subjects would no longer be able to distinguish spectral differences between consonants that have the same manner and voicing but differ in the place they are produced in the mouth (e.g., /p/ [labial] vs. /t/ [alveolar] vs. /k/ [velar]). To verify this, confusion matrices for each test condition for each individual subject were submitted to a feature analysis. The features entered into the analysis were voicing (voiced/unvoiced), manner (stop/fricative/nasal) and place (front/middle/back). The means and standard deviations across the ten subjects are shown in Table 5. The small standard deviations indicate that error patterns were very similar across individuals. As expected, conversion to a small number of channels almost completely removed the place information which would be coded in the signal fine structure. Voicing and manner, however, were well preserved with as few as two channels, suggesting that these types of information are encoded by the modulation envelope.

Discussion

The goal of this project was to examine the extent to which the average modulation spectrum of individual phonemes can be used to predict the confusions that human listeners will experience when presented with stimuli in which the carrier information is largely removed. In order to address this question in the most general terms, a relatively simple model was tested, using only six carrier bands and six modulation filters. In addition, the time-varying nature of the output of the modulation filters was suppressed by summing the energy passed by a given filter across the duration of the individual phonemes. Given this very basic modeling exercise, it is informative to note how strong the correlations are between human performance and similarity of modulation spectra as assessed by the spectral cross-correlation. In all five of the conditions tested (one-band, two-band, four-band, eight-band and unprocessed), the correlations are positive and significant. This result suggests that even such a simplified measure of the modulation information in speech can be useful for predicting human performance and suggests that modulation sensitivity is one of the mechanisms underlying that performance.

The focus of this study was the information conveyed by signal modulations at multiple rates, beyond that conveyed by the carrier signal. To evaluate the importance of the modulation spectrum irrespective of carrier frequency, it was necessary to select a processing method that would retain modulation spectrum cues but degrade the carrier frequency information. The signals used here provided varying degrees of carrier frequency information, ranging from none (one-channel) to eight channels. Because fine-structure cues were eliminated by the

processing, even the eight-channel signal provided only gross spectral information. We can evaluate the success of this processing choice by viewing the feature analysis. Given the absence of fine-structure in these signals (and our intent to restrict carrier frequency cues), we would expect poor transmission of place information; particularly with a small number of channels. Voicing should still be available, either from the modulation spectrum itself (Christiansen & Greenberg, 2005) or via periodicity information preserved by the processing. Likewise, manner information should be transmitted by the modulation spectrum (Christiansen & Greenberg, 2005; van der Horst, Leeuw, & Dreschler, 1999). All of these expectations are supported by the feature analysis (Table 5).

It is interesting to think about Figure 7 as a measure of modulation *dissimilarity*. That is, when the modulation spectrum of one phoneme is unlike another, those two phonemes are never confused (i.e, there are no points on the graph representing low SCI/high error rates). But listeners only sometimes confuse phonemes that have similar modulation spectra (i.e., the points on the graph representing high SCI/high error rates). This probably reflects information that is available to the listener but not reflected in the simplified modulation spectrum. As discussed above, fine-structure temporal cues would be eliminated by the processing. However, beginning with the two-channel signal, rudimentary spectral cues are available to the listeners that are not reflected in the SCI, due to the use of normalized modulation index values. Also not reflected in our measure, which represents modulation energy only through 32 Hz, are periodicity cues (cf. Rosen, 1992). Such cues are the likely reason why not every high SCI results in errors. Nonetheless, it is strong support for the general approach of the model that as the number of channels increases, the accuracy of the error prediction increases as well. This is true despite the reduction in the number of points available for the analysis (due to the reduced number of combinations producing at least 5% errors).

An additional aspect of the analysis concerns the degree to which the modulation spectrum varies as a function of carrier frequency. Consistent with Crouzet and Ainsworth (2001), it was found that modulation spectrum does indeed vary across carrier frequency (see Fig. 3) but Fig. 3 also demonstrates the general finding that modulation information is highly correlated across bands. This is consistent with data from Apoux and Bacon (2004) that suggest that for consonant identification in quiet, any one of four carrier bands can be removed without a critical drop in performance; and from Christiansen and Greenberg (2005), who showed that removal of high-frequency modulations within one of three spectral slits had little effect on transmitted consonant information.

Implications for Future Work

Having shown the usefulness of this simple and straightforward measure of modulation spectrum (based on a set of thirty-six values), it will be important in the future to evaluate the additional gain in prediction ability that comes from adopting more complex measures. Two areas in which substantial work has already been done are the carrier-frequency filtering of the peripheral auditory system and the extraction of complex modulation waveforms rather than simply modulation energy. What is not yet known, however, is the degree to which a better fit to the human cochlear filtering data or a more faithful representation of the signal processing necessary for accurately reconstructing speech waveforms will lead to better predictions of human performance. Gallun and Hafter (2006), for example, obtained quite accurate predictions of human listeners' ability to detect changes in the intensity of ongoing tones in the presence of on and off-frequency modulated maskers by using a model of modulation sensitivity with less resolution in carrier frequency than the one used here. In fact, it was essential to making accurate predictions that the filtering in carrier frequency was *substantially broader* than that found in models of cochlear physiology. Indeed, nearly all psychophysical investigations of modulation sensitivity that have examined off-frequency modulation masking

(e.g., Yost and Sheft, 1989) have shown interactions that would be impossible given the auditory filters measured for the detection of tones in noise.

In future investigations, it will also be important to extend the analysis to speech tokens of longer duration than those studied here. Substantial similarity in the duration of these stimuli may actually have made differences in modulation spectrum more difficult to detect due to the introduction of peaks in the modulation spectrum at the frequency corresponding the syllable duration (250 ms, or 4 Hz). Similarly, the analysis used here assumed that the information at all frequencies was equally important. It is likely that better results could be obtained by differentially weighting the information at some frequencies relative to others. The relative importance of modulation across carrier bands may differ as well, as suggested by previous data showing that modulations in high-frequency carrier bands may be more important, at least for speech in background noise (Apoux & Bacon, 2004).

In terms of extending the analysis, it will be important to address the issue of individual talker variability. All of the analyses presented in this paper were for the combined data from four talkers. This choice was reasonable given that the listeners were always presented with speech tokens that were drawn randomly from these four talkers. However, the modulation spectrum of a given phoneme is not invariant across talkers in the representation used in this paper. In addition, some talkers had spectra that were quite similar across phonemes, while others differed substantially. Even the features that seemed to distinguish two phonemes for one talker were not necessarily present for another. Although the model (and the listeners) seemed to handle this variability fairly well, it will be worth examining this aspect of the modulation spectrum more carefully in the future.

Finally, there are a number of reasons to believe that listeners are sensitive to ongoing changes in modulation as well as time-averaged modulation energy. While it must necessarily complicate the simple model used in this paper, the actual information used by human listeners is probably better represented by thirty-six time-varying waveforms than by thirty-six static values. Were it feasible to characterize the changes in the amplitude and phase of each modulation channel over time, rather than assigning a single static value to each by averaging across time, a more complete representation of the modulation information would be available. Imagine, for example, being able to make use of the fact the fact that 4 Hz modulation in the highest carrier frequency reached a maximum slightly before the 4 Hz modulation in the lowest carrier frequency. Such a representation could capture glides across carrier frequency at various rates as well as fine timing differences in the onsets of energy in different modulation bands and/or carrier bands. In the future, the challenge will be to characterize the phase and timing information in the modulation filter outputs in such a way that human performance can be easily and accurately predicted. For now, it seems that the method described in this paper is quite successful in maximizing predictive power while minimizing the amount of information required by the model.

Acknowledgments

The research reported here was supported by the Department of Veterans Affairs, Veterans Health Administration, Rehabilitation Research and Development Service (F.G.), grants DC006014 and DC04661 from the National Institute for Deafness and Communication Disorders (P.S.), and by the Bloedel Hearing Research Center (P.S.). The authors thank Stuart Rosen and the University College London Department of Phonetics and Linguistics for providing the FIX program, and Eric Hoover for his help with data collection. The authors are also grateful to Dr. Wesley Grantham and two anonymous reviewers for their comments during the review process.

Appendix A

Confusion matrix for 1-channel condition. Each stimulus was identified eighty times (each of the four talkers' utterances was identified twice, resulting in eight identifications by each of ten different listeners for each phoneme).

		Stimulus									
		/aða/	**/aba/**	**/ada/**	**/afa/**	**/aga/**	**/aka/**	**/ama/**	**/ana/**	**/apa/**	**/asa/**
Response	/aða/	15	10	8	5	5	2	3	4	0	6
	/aba/	1	25	22	9	10	4	1	4	1	2
	/ada/	4	2	17	5	13	0	4	1	3	1
	/afa/	2	5	1	15	1	3	0	3	0	24
	/aga/	2	1	3	0	0	3	1	1	1	0
	/aka/	0	1	1	3	6	17	1	1	12	0
	/ama/	1	4	2	0	2	1	15	10	1	0
	/ana/	0	4	0	0	0	0	5	9	1	0
	/apa/	0	3	0	5	3	1	0	1	12	3
	/asa/	4	3	2	11	4	1	7	8	3	27
	/aʃa/	1	1	0	1	1	0	11	7	0	0
	/ata/	1	5	7	8	10	44	1	1	44	1
	/aθa/	1	6	4	16	7	1	3	2	1	14
	/ava/	33	8	13	1	18	0	18	18	1	0
	/aza/	15	1	0	1	0	0	7	7	0	2
	/aʒa/	0	1	0	0	0	3	3	3	0	0

Appendix B

Confusion matrix for 2-channel condition. Each stimulus was identified eighty times (each of the four talkers' utterances was identified twice, resulting in eight identifications by each of ten different listeners for each phoneme).

		Stimulus									
		/aða/	**/aba/**	**/ada/**	**/afa/**	**/aga/**	**/aka/**	**/ama/**	**/ana/**	**/apa/**	**/asa/**
Response	/aða/	3	0	0	3	0	0	4	1	0	0
	/aba/	11	71	63	3	48	0	9	9	2	0
	/ada/	1	4	12	0	8	0	0	1	1	0
	/afa/	1	1	0	52	0	0	0	0	0	57
	/aga/	5	1	3	0	12	0	0	3	2	0
	/aka/	1	1	0	0	2	53	0	0	18	0
	/ama/	0	0	0	0	0	0	40	25	0	0
	/ana/	1	0	0	0	0	0	15	29	0	0
	/apa/	0	1	1	1	5	13	0	0	51	0
	/asa/	0	0	0	1	1	0	0	0	0	8
	/aʃa/	0	0	0	0	0	0	0	0	0	8
	/ata/	0	0	0	0	0	14	0	0	6	0

		Stimulus									
		/aða/	/aba/	/ada/	/afa/	/aga/	/aka/	/ama/	/ana/	/apa/	/asa/
	/aθa/	0	1	0	17	0	0	1	1	0	7
	/ava/	56	0	1	2	4	0	10	10	0	0
	/aza/	1	0	0	1	0	0	1	1	0	0
	/aʒa/	0	0	0	0	0	0	0	0	0	0

Appendix C

Confusion matrix for 4-channel condition. Each stimulus was identified eighty times (each of the four talkers' utterances was identified twice, resulting in eight identifications by each often different listeners for each phoneme).

		Stimulus									
		/aða/	/aba/	/ada/	/afa/	/aga/	/aka/	/ama/	/ana/	/apa/	/asa/
Response	/aða/	11	0	1	3	0	0	0	0	0	0
	/aba/	5	75	48	1	24	0	4	9	0	0
	/ada/	3	1	22	0	12	0	0	0	0	0
	/afa/	1	0	0	50	0	0	0	0	1	10
	/aga/	4	1	8	1	41	0	0	1	0	0
	/aka/	0	0	0	0	0	58	0	1	6	0
	/ama/	0	2	0	0	0	0	65	41	0	0
	/ana/	0	0	0	0	0	0	7	24	0	0
	/apa/	0	1	1	0	3	3	0	0	61	0
	/asa/	0	0	0	7	0	0	0	0	0	42
	/aʃa/	0	0	0	0	0	0	0	0	0	15
	/ata/	0	0	0	0	0	19	0	0	12	0
	/aθa/	2	0	0	18	0	0	0	2	0	9
	/ava/	52	0	0	0	0	0	3	2	0	2
	/aza/	2	0	0	0	0	0	0	0	0	2
	/aʒa/	0	0	0	0	0	0	1	0	0	0

Appendix D

Confusion matrix for 8-channel condition. Each stimulus was identified eighty times (each of the four talkers' utterances was identified twice, resulting in eight identifications by each often different listeners for each phoneme).

		Stimulus									
		/aða/	/aba/	/ada/	/afa/	/aga/	/aka/	/ama/	/ana/	/apa/	/asa/
Response	/aða/	37	0	0	0	0	0	0	0	0	0
	/aba/	1	76	9	0	0	0	0	1	0	0
	/ada/	3	0	63	0	3	0	0	1	0	0
	/afa/	0	1	0	55	0	1	0	0	0	6

		Stimulus									
		/aða/	/aba/	/ada/	/afa/	/aga/	/aka/	/ama/	/ana/	/apa/	/asa/
	/aga/	8	1	7	0	76	1	0	1	1	0
	/aka/	0	0	0	0	1	72	0	0	3	0
	/ama/	0	0	0	0	0	0	75	19	0	0
	/ana/	0	0	0	0	0	0	4	58	0	0
	/apa/	0	2	0	0	0	0	0	0	75	0
	/asa/	0	0	0	8	0	0	0	0	0	65
	/aʃa/	0	0	0	0	0	0	0	0	0	2
	/ata/	0	0	0	0	0	6	0	0	1	0
	/aθa/	5	0	0	16	0	0	0	0	0	5
	/ava/	21	0	1	1	0	0	1	0	0	0
	/aza/	5	0	0	0	0	0	0	0	0	2
	/aʒa/	0	0	0	0	0	0	0	0	0	0

Appendix E

Confusion matrix for unprocessed condition. Each stimulus was identified eighty times (each of the four talkers' utterances was identified twice, resulting in eight identifications by each often different listeners for each phoneme).

		Stimulus									
		/aða/	/aba/	/ada/	/afa/	/aga/	/aka/	/ama/	/ana/	/apa/	/asa/
Response	/aða/	68	1	0	0	0	0	0	0	0	1
	/aba/	0	78	0	0	0	0	0	0	0	0
	/ada/	0	1	80	0	0	0	0	0	0	0
	/afa/	0	0	0	63	0	0	0	0	0	0
	/aga/	0	0	0	0	80	0	0	0	0	0
	/aka/	0	0	0	0	0	80	0	0	0	0
	/ama/	0	0	0	0	0	0	80	0	0	0
	/ana/	0	0	0	0	0	0	0	79	0	0
	/apa/	0	0	0	0	0	0	0	0	80	0
	/asa/	0	0	0	5	0	0	0	0	0	79
	/aʃa/	0	0	0	0	0	0	0	0	0	0
	/ata/	0	0	0	0	0	0	0	1	0	0
	/aθa/	6	0	0	12	0	0	0	0	0	0
	/ava/	4	0	0	0	0	0	0	0	0	0
	/aza/	2	0	0	0	0	0	0	0	0	0
	/aʒa/	0	0	0	0	0	0	0	0	0	0

References

American National Standards Institute. Methods for calculation of the speech intelligibility index (ANSI 3.5-1997). ANSI; New York: 1997.

American National Standards Institute. Specifications for Audiometers (ANSI 3.6-2004). ANSI; New York: 2004.

Apoux F, Bacon SP. Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. Journal of the Acoustical Society of America 2004;116 (3):1671–1680. [PubMed: 15478433]

Arai T, Pavel M, Hermansky H, Avendano C. Syllable intelligibility for temporally filtered LPC cepstral trajectories. Journal of the Acoustical Society of America 1999;105(5):2783–2791. [PubMed: 10335630]

Arai, T.; Greenberg, S. The temporal properties of spoken Japanese are similar to those of English. Proceedings of Eurospeech-97; Rhodes. 1997. p. 1011-1014.

Arai, T.; Greenberg, S. Speech intelligibility in the presence of cross-channel spectral asynchrony. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing; Seattle. 1998. p. 933-936.

Atlas, L.; Li, Q.; Thompson, J. Homomorphic modulation spectra. Proceedings of ICASSP; Montreal, Quebec, Canada. 2004. p. 761-764.

Bacon SP, Grantham DW. Modulation masking: Effects of modulation frequency, depth, and phase. Journal of the Acoustical Society of America 1989;85(6):2575–2580. [PubMed: 2745880]

Chi T, Gao Y, Guyton MC, Ru P, Shamma S. Spectro-temporal modulation transfer functions and speech intelligibility. Journal of the Acoustical Society of America 1999;106(5):2719–2732. [PubMed: 10573888]

Chi T, Ru P, Shamma S. Multiresolution spectrotemporal analysis of complex sounds. Journal of the Acoustical Society of America 2005;118(2):887–906. [PubMed: 16158645]

Christiansen, TE.; Greenberg, S. Frequency selective filtering of the modulation spectrum and its impact on consonant identification. 21st Danavox Symposium “Hearing Aid Fitting”; Kolding, Denamrak. 2005.

Cox RM, Moore JN. Composite speech spectrum for hearing aid gain prescriptions. Journal of Speech and Hearing Research 1988;31(1):102–107. [PubMed: 3352247]

Crouzet, O.; Ainsworth, WA. Envelope information in speech processing: Acoustic-phonetic analysis vs. auditory figure-ground segregation. Proceedings of the 7th European Conference on Speech Communication and Technology; Aalborg, Denmark. 2001. p. 477-480.

Dau T, Kollmeier B, Kohlrausch A. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. Journal of the Acoustical Society of America 1997a;102(5): 2892–2905. [PubMed: 9373976]

Dau T, Kollmeier B, Kohlrausch A. Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. Journal of the Acoustical Society of America 1997b;102(5):2906–2919. [PubMed: 9373977]

Drullman R, Festen JM, Plomp R. Effect of temporal envelope smearing on speech reception. Journal of the Acoustical Society of America 1994a;95(2):1053–1064. [PubMed: 8132899]

Drullman R, Festen JM, Plomp R. Effect of reducing slow temporal modulations on speech reception. Journal of the Acoustical Society of America 1994b;95(5):2670–2680. [PubMed: 8207140]

Dudley H. Remaking speech. Journal of the Acoustical Society of America 1939;11(2):169–177.

Ewert SD, Dau T. Characterizing frequency selectivity for envelope fluctuations. Journal of the Acoustical Society of Americ 2000;108:1181–1196.

Fletcher H. Auditory patterns. Reviews of Modern Physics 1940;12:47–66.

Gallun FJ, Hafter ER. Amplitude modulation as a mechanism for increment detection. Journal of the Acoustical Society of America 2006;119(6):3919–3930. [PubMed: 16838535]

Green T, Katiri S, Faulkner A, Rosen S. Talker intelligibility differences in cochlear implant listeners. Journal of the Acoustical Society of America 2007;121(6):EL223–229. [PubMed: 17552573]

Greenberg, S.; Arai, T. The relation between speech intelligibility and the complex modulation spectrum. Proceedings of the 7th European Conference on Speech Communication and Technology; Aalborg, Denmark. 2001. p. 473-476.

Greenberg, S.; Arai, T.; Silipo, R. Speech intelligibility derived from exceedingly sparse spectral information. Proceedings of the International Conference of Spoken Language Processing; Sydney, Australia. 1998. p. 2803-2806.

Greenberg S, Kingsbury BED. The modulation spectrogram: In pursuit of an invariant representation of speech. ICASSP 1997 1997:1647–1650.

Hermansky, H. The modulation spectrum in the automatic recognition of speech. Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding; Santa Barbara, California. 1997. p. 140-147.

Houtgast T. Frequency selectivity in amplitude-modulation detection. Journal of the Acoustical Society of America 1989;85(4):1676–1680. [PubMed: 2708683]

Kanedera N, Arai T, Hermansky H, Pavel M. Effect of reducing slow temporal modulations on speech reception. Speech Communication 1999;28:43–55.

Krause JC, Braida LD. Acoustic properties of naturally produced clear speech at normal speaking rates. Journal of the Acoustical Society of America 2004;115(1):362–378. [PubMed: 14759028]

Mueller HG, Killion MC. An easy method for calculating the Articulation Index. Hearing Journal 1990;9:14–17.

Pittman AL, Stelmachowicz PG, Lewis DE, Hoover BM. Spectral characteristics of speech at the ear: Implications for amplification in children. Journal of Speech, Language and Hearing Research 2003;46(3):649–657.

Rosen S. Temporal information in speech: Acoustic, auditory and linguistic aspects. Philosophical Transactions: Biological Sciences 1992;336(1278):367–373. [PubMed: 1354376]

Schroeder MR. Reference signal for signal quality studies. Journal of the Acoustical Society of America 1968;44(6):1735–1736.

Scollie S, Seewald R, Cornelisse L, Moodie S, Bagatto M, Laurnagarav D, et al. The Desired Sensation Level multistage input/output algorithm. Trends in Amplification 2005;9(4):159–197. [PubMed: 16424945]

Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science 1995;270(5234):303–304. [PubMed: 7569981]

Silipo, R.; Greenberg, S.; Arai, T. Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations. Proceedings of the 6th European Conference on Speech Communication and Technology; 1999. p. 2687-2690.

van der Horst R, Leeuw AR, Dreschler WA. Importance of temporal-envelope cues in consonant recognition. Journal of the Acoustical Society of America 1999;105(3):1801–1809. [PubMed: 10089603]

Vinton, MS.; Atlas, LE. A scalable and progressive audio codec; Paper presented at the ICASSP; 2001; 2001.

Yost WA, Sheft S. Across-critical-band processing of amplitude-modulated tones. Journal of the Acoustical Society of America 1989;85(2):848–857. [PubMed: 2925999]

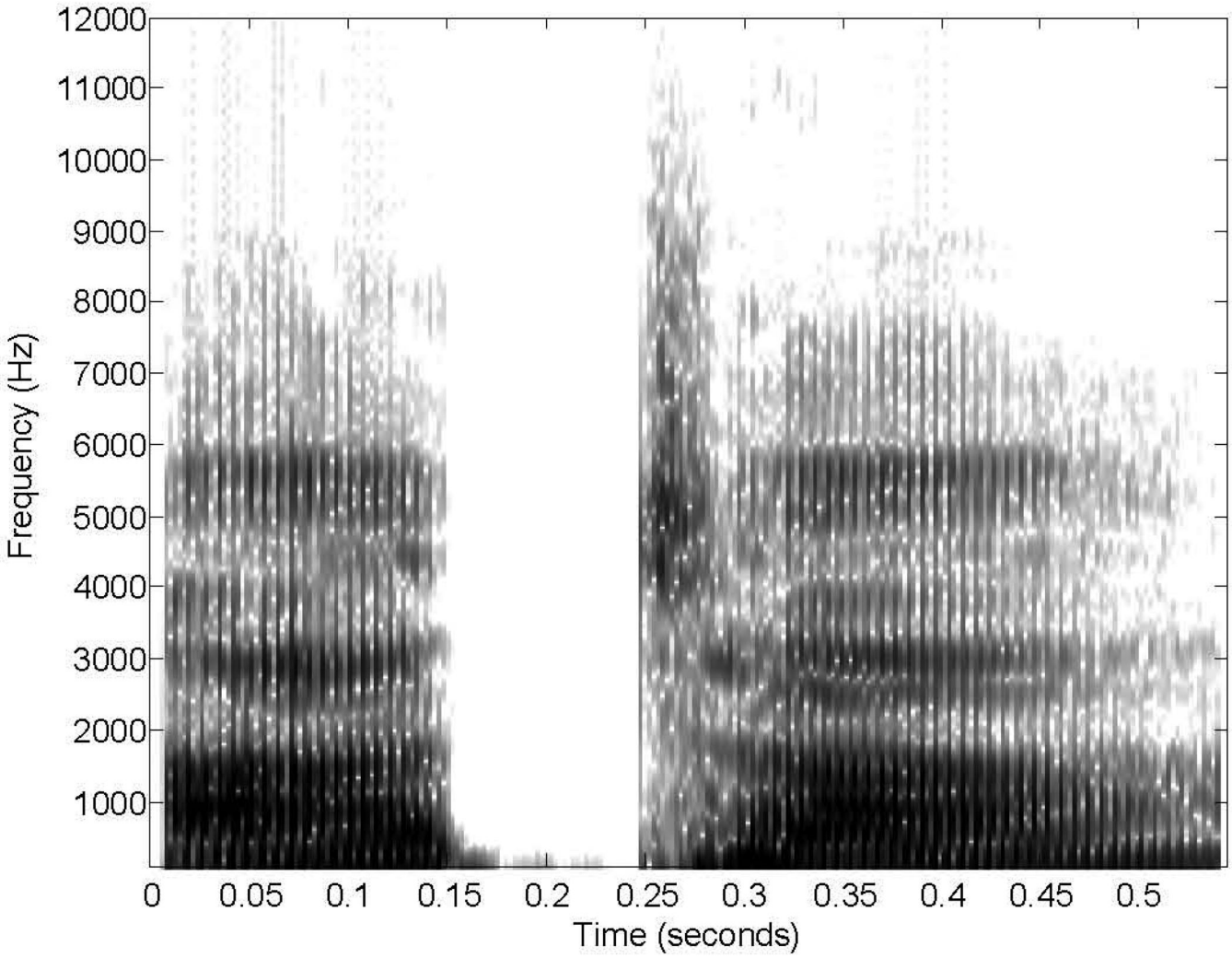

Figure 1.
Spectrogram of the phoneme /ata/ with spectral energy plotted as a function of time.

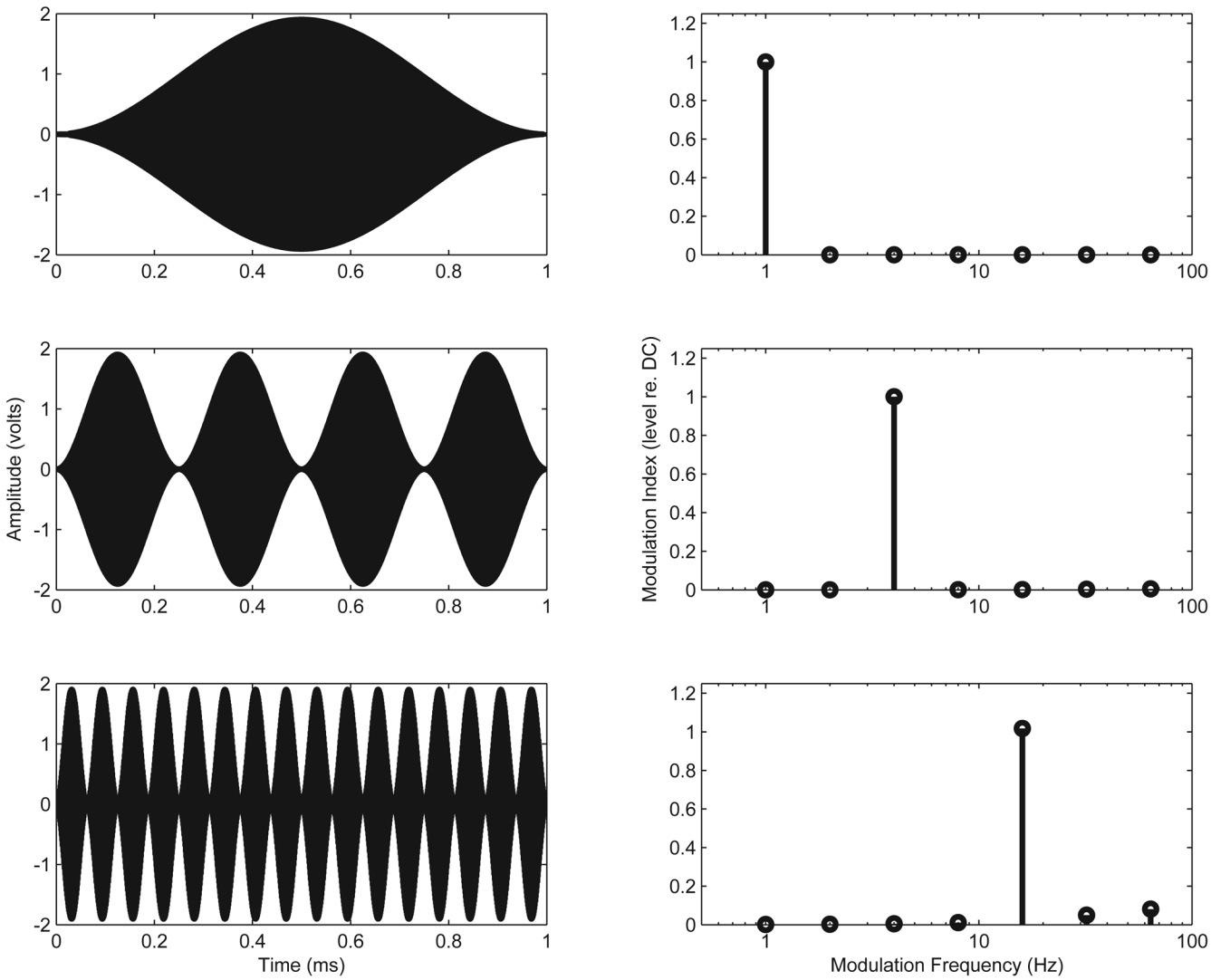

Figure 2.
Examples of the modulation spectra for sinusoidal amplitude modulation (SAM) imposed on a 1000 Hz carrier. Time waveform is on the left and the modulation spectra (re. DC) is on the right. Top panels:1 Hz SAM; middle panels: 4 Hz SAM; bottom panels:16 Hz SAM. See text for details of the analysis.

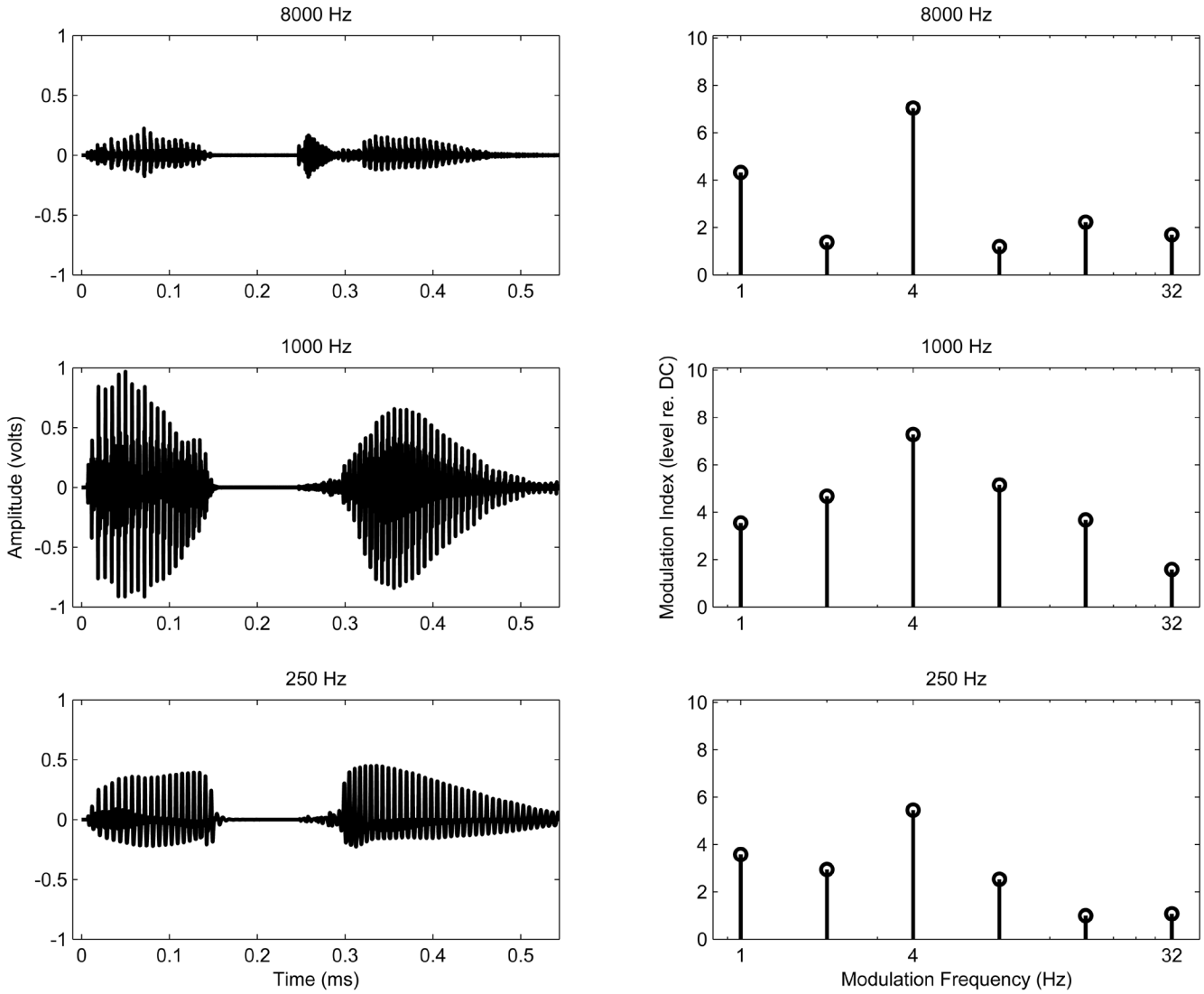

Figure 3.

Modulation spectra analyzed for three octave-wide bands of the phoneme /ata/ shown in Figure 1. Time waveform is on the left and the modulation spectra (relative to the energy at 0 Hz or "DC") is on the right. Top panels: 8000 Hz center frequency (cf); middle panels: 1000 Hz cf; bottom panels: 250 Hz cf. See text for details of the analysis.

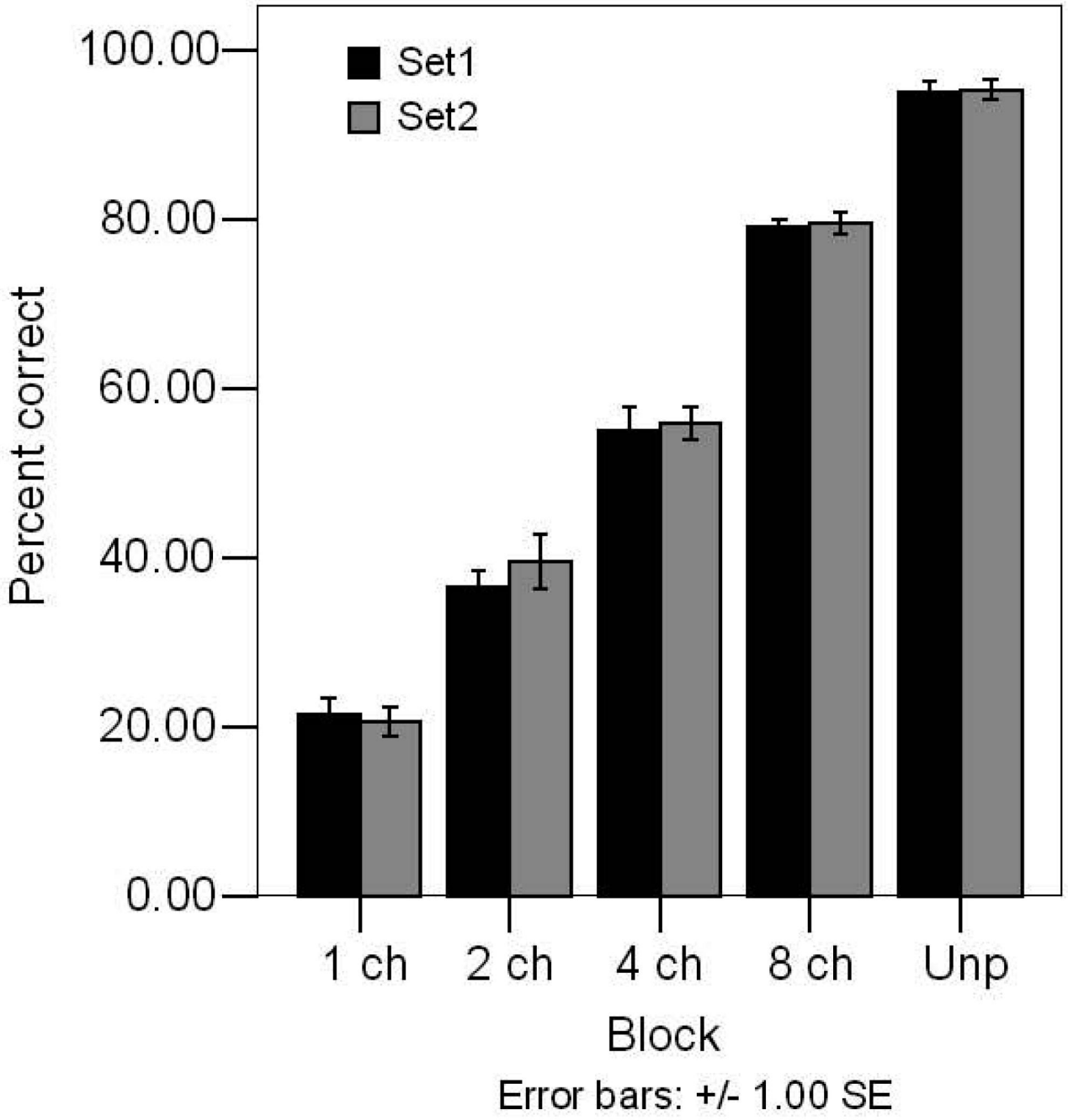

Figure 4.
Mean scores for the first (black bars) and second (grey bars) set for each of the five test conditions, across ten listeners. Error bars represent +/- one standard error about the mean.

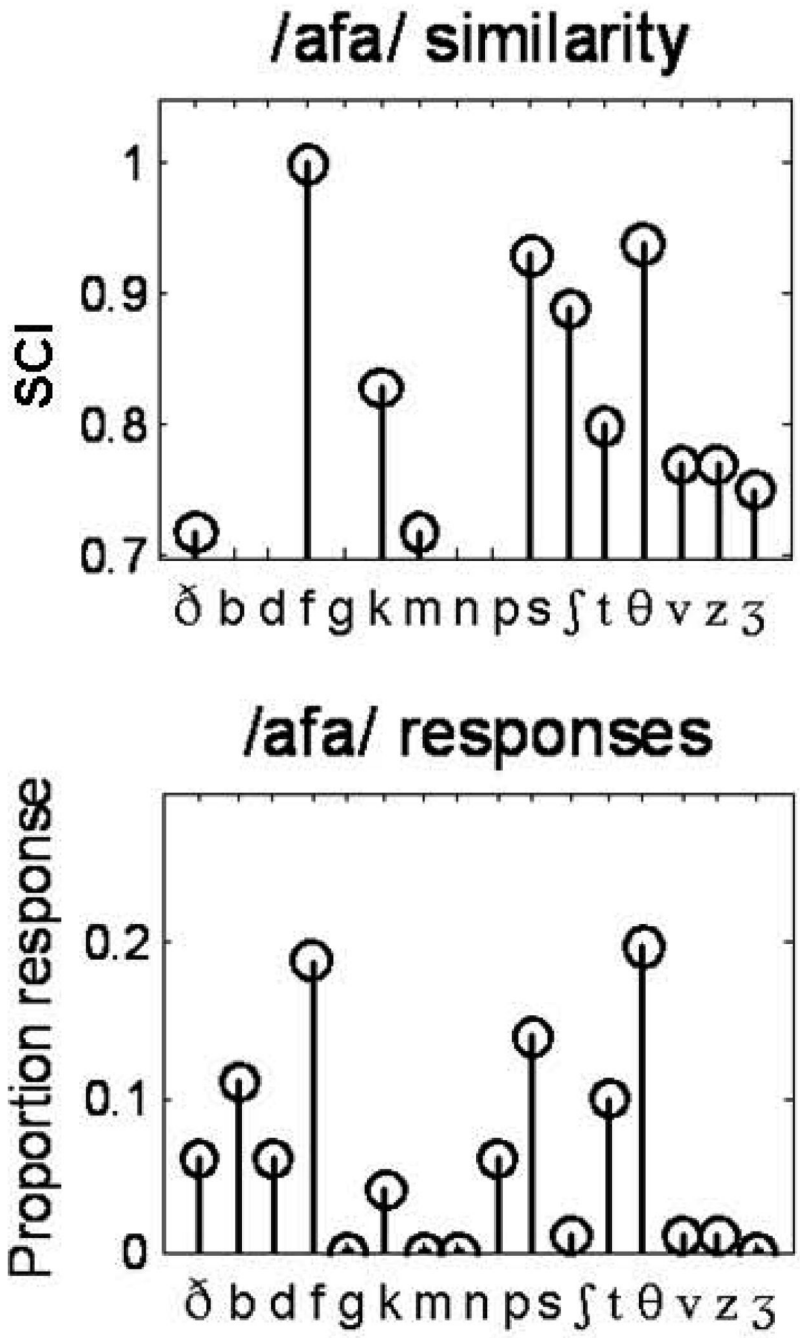

Figure 5.
Modulation spectrum similarity and response pattern for /afa/ in the one-channel SCN condition. The top "similarity" panel shows the spectral cross-correlation between /afa/ and the 15 other phonemes. /afa/ has a value of 1 with itself. The bottom panel shows the proportion of times /afa/ was chosen.

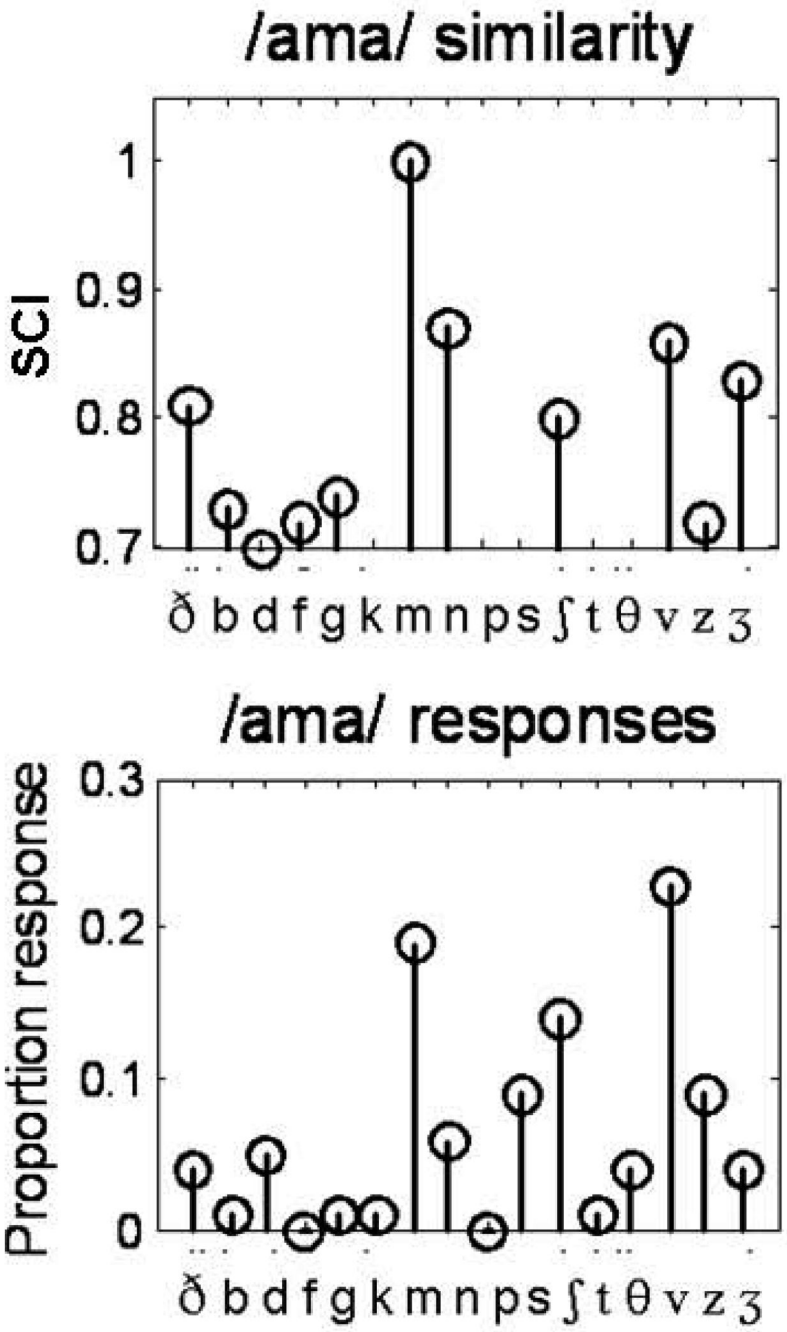

Figure 6.
Modulation spectrum similarity and response pattern for /ama/ in the one-channel SCN condition. The top "similarity" panel shows the spectral cross-correlation between /ama/ and the 15 other phonemes. /ama/ has a value of 1 with itself. The bottom panel shows the proportion of times /ama/ was chosen.

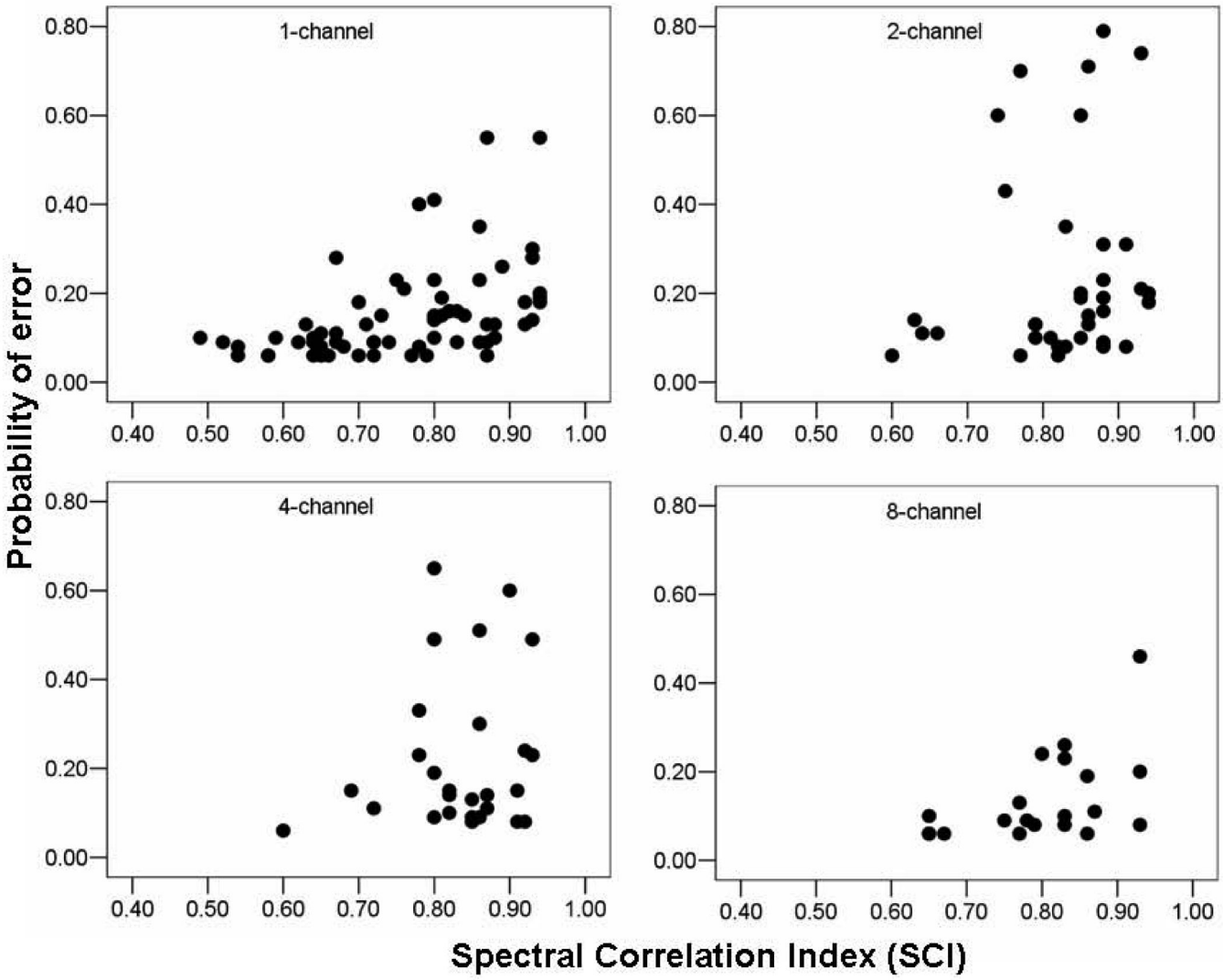

Figure 7.

Modulation spectrum similarity of an incorrectly identified phoneme to the presented phoneme (the SCI value) and the proportion of the trials on which the presented phoneme resulted in that specific error. Data do not include correct responses, or phonemes incorrectly identified less than 5% of the time. Each panel shows one of the processed test conditions, with data drawn from the master confusion matrices presented in Appendices A-D.

Table 1

Crossover frequencies used for filtering.

Condition	Crossover frequencies
one-channel	n/a
two-channel	1130 Hz
four-channel	440 Hz, 1130 Hz, 2800 Hz
eight-channel	280 Hz, 440 Hz, 710 Hz, 1130 Hz, 1780 Hz, 2800 Hz, 4440 Hz

Table 2

Spectral correlation index (SCI) values indicating modulation similarity for the five types of processing used. Each value represents the correlation of two 2,304 point vectors (36 modulation index values for each of 16 phonemes for each of 4 talkers).

Condition	two-channel	four-channel	eight-channel	unprocessed
one-channel	0.921	0.900	0.873	0.871
two-channel		0.923	0.886	0.874
four-channel			0.939	0.938
eight-channel				0.961

Table 3

Spectral correlation index (SCI) values indicating modulation similarity for the unprocessed phonemes, calculated for spectra combined across talkers. Each value represents the correlation of two 144 point vectors (36 modulation indexes for each of 4 talkers).

	/aba/	**/ada/**	**/afa/**	**/aga/**	**/aka/**	**/ama/**	**/ana/**	**/apa/**	**/asa/**	**/aʃa/**	**/ata/**	**/aθa/**	**/ava/**	**/aza/**	**/aʒa/**
/aða/	0.70	0.62	0.72	0.64	0.71	0.82	0.84	0.56	0.66	0.68	0.67	0.66	0.85	0.81	0.78
/aba/		0.92	0.67	0.88	0.83	0.75	0.74	0.85	0.56	0.62	0.84	0.59	0.88	0.59	0.66
/ada/			0.67	0.82	0.84	0.73	0.71	0.85	0.58	0.57	0.88	0.60	0.80	0.58	0.60
/afa/				0.66	0.82	0.81	0.75	0.74	0.88	0.77	0.82	0.96	0.77	0.79	0.70
/aga/					0.70	0.73	0.74	0.83	0.52	0.62	0.76	0.59	0.80	0.53	0.69
/aka/						0.75	0.66	0.87	0.81	0.70	0.95	0.77	0.75	0.73	0.68
/ama/							0.88	0.69	0.71	0.69	0.78	0.79	0.84	0.81	0.72
/ana/								0.63	0.65	0.63	0.70	0.72	0.83	0.76	0.73
/apa/									0.67	0.59	0.90	0.69	0.72	0.57	0.59
/asa/										0.81	0.79	0.90	0.64	0.84	0.71
/aʃa/											0.69	0.75	0.68	0.76	0.87
/ata/												0.79	0.76	0.69	0.66
/aθa/													0.69	0.77	0.65
/ava/														0.71	0.70
/aza/															0.81

Table 4

Bivariate correlations between modulation similarity and probability of making a specific error, calculated across all presented phonemes for all listeners by using the master confusion matrices.

Condition	one-channel	two-channel	four-channel	eight-channel	Unprocessed
Pearson *r*	.453	.343	.570	.816	.739
p	<.001	.015	<.001	<.001	<.001
N	80	50	43	36	22

Table 5

Results of feature analysis (values in proportion of information transmitted). Values show mean (and standard deviation) across ten subjects.

Condition	Voicing	Manner	Place
One-channel	.30 (.16)	.35 (.10)	.06 (.05)
Two-channel	.77 (.14)	.73 (.08)	.14 (.05)
Four-channel	.81 (.12)	.82 (.08)	.29 (.07)
Eight-channel	.92 (.07)	.91 (.06)	.54 (.07)
Unprocessed	.94 (.09)	.99 (.02)	.90 (.03)