Properties of Naturally Produced Clear Speech at Normal Rates and Implications for Intelligibility Enhancement

by

Jean Christine Krause

B.S.E.E., Georgia Institute of Technology (1993)
S.M., Massachusetts Institute of Technology (1995)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2001

Author .....
Department of Electrical Engineering and Computer Science
February 5, 2001

Certified by .....
Louis D. Braida
Henry E. Warren Professor of Electrical Engineering
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Properties of Naturally Produced Clear Speech at Normal Rates and Implications for Intelligibility Enhancement

by

Jean Christine Krause

Submitted to the Department of Electrical Engineering and Computer Science
on February 5, 2001, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In adverse listening conditions, talkers can increase their intelligibility by speaking clearly[45, 55, 42]. While producing clear speech, however, talkers often reduce their speaking rate significantly[46, 55]. A recent study[28] showed that speaking slowly is not responsible for the high intelligibility of clear speech, since talkers can produce clear speech at normal rates with training. This finding suggests that acoustical factors other than reduced speaking rate are responsible for the high intelligibility of clear speech. To gain insight into these factors, acoustical properties (global, phonological, and phonetic) of conversational and clear speech produced at normal speaking rates were examined. Three global acoustic properties associated with clear/normal speech were identified: increased energy near the second and third formants, higher average and greater range of F0, and increased modulation depth of low frequency modulations of the intensity envelope. In order to determine which of these acoustical properties of clear/normal speech contribute most to its high intelligibility, signal processing transformations of conversational speech were developed. Results of intelligibility tests with hearing-impaired listeners and normal hearing listeners in noise suggest that these properties may not fully account for the intelligibility benefit of clear/normal speech. Other properties important for highly intelligible speech may not have been identified in this study due to the complexity of the acoustic database and varying talker strategies.

Thesis Supervisor: Louis D. Braida
Title: Henry E. Warren Professor of Electrical Engineering

Acknowledgments

Financial support for this work was provided by a grant from the National Institute on Deafness and Other Communication Disorders and a National Defense Science and Engineering Graduate fellowship from the Office of Naval Research.

I would like to thank my thesis commmittee, Louis Braida, Joe Perkell, Jae Lim, and Ken Stevens. Their technical expertise, advice, and comments on drafts greatly strengthened the final product. In particular, I am grateful to Louis Braida, who served as thesis supervisor and hence was most involved in keeping the project moving along, providing not only many technical suggestions but also encouragement over the years.

In addition to the thesis committee, there are a number of people that deserve thanks for their direct participation in the details of the work. First, I am grateful to the many listeners who participated in the various intelligibility experiments, most especially my father, who flew a thousand miles to participate "just because." The listeners' task was always tedious and often difficult, and yet each of them bore it good-naturedly. Second, UROPs Matt and Debbie were of enormous assistance in phonetically labelling the database. Third, several people helped with scoring sentences from the numerous intelligibility experiments, most notably Melissa, Cathy, and Jay.

Support from the people in the Sensory Communications Group came in a variety of forms: lively technical (and non-technical!) discussions, shared MATLAB utilities and other computer/software/hardware support, lunch room banter, late-night pep talks, chocolate (lots of chocolate!), good friendships, and general camaraderie.

Finally, I thank my family for being a constant source of love, support, and encouragement through it all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When confronted with difficult communication environments, many speakers adopt a speaking style which permits them to be understood more easily. Previous studies have demonstrated that this speaking style, known as *clear speech*, is significantly more intelligible than conversational speech for both hearing-impaired listeners in a quiet background[45, 55] and normal hearing listeners in noise[55, 42, 28] as well as for normal hearing and hearing-impaired listeners in noise and reverberation backgrounds[42]. Furthermore, the intelligibility advantage is independent of listener, presentation level, and frequency-gain characteristic[45]. These results suggest that signal processing schemes that convert conversational speech to a sufficiently close approximation of clear speech could improve speech intelligibility in many situations. Moreover, the intelligibility improvement provided by these signal processing schemes should be independent of any benefits obtained from varying the frequency-gain characteristic of the linear amplification found in conventional hearing aids.

In order to implement such signal processing schemes, however, it is first necessary to identify the acoustical factors responsible for the high intelligibility of clear speech. While many acoustical differences between clear speech and conversational speech have been described[46], the specific characteristics of clear speech responsible for its high intelligibility have not yet been isolated. Identifying these characteristics has proven difficult, since the effect of speaking rate on intelligibility has not been well understood until recently. In particular, it had not been determined whether

the high intelligibility of clear speech resulted primarily from its particular acoustical characteristics or simply from the reduction in speaking rate typically exhibited in clear speech. This question is particularly important for real-time hearing aid applications, since audio and visual signals must remain synchronized for maximum benefit to the listener.

A recent study[28], however, has shown that reduced speaking rate is not necessary for the high intelligibility of clear speech, demonstrating that with training, talkers can produce clear speech at normal speaking rates. This finding suggests that acoustical factors other than speaking rate can be responsible for the intelligibility advantage of clear speech over conversational speech. This thesis attempts to identify some of these acoustical factors.

Chapter 2

Background

Research on naturally produced clear speech dates back several decades. In recent years, however, the focus of this research has shifted from investigating intertalker differences to investigating intratalker differences between clear and conversational speech. In an effort to identify the essential components of highly intelligible speech, these studies have investigated the properties of clear speech as well as signal processing schemes for intelligibility enhancement of conversational speech. This chapter describes these intratalker studies as well as other relevant attempts at intelligibility enhancement.

2.1 Properties of Clear Speech

Previous reports describe intelligibility differences, acoustical differences, and speaking rate differences between conversational and clear speech. Identifying the specific acoustic characteristics of clear speech responsible for its high intelligibility has proven difficult, however, since the effects of reduced speaking rate on the intelligibility of clear speech were not well understood. In particular, the extent to which the high intelligibility of clear speech resulted from a reduced speaking rate had not been determined. Therefore, recent work by Krause[28], has been concerned with characterizing the effects of changes in speaking rate on the intelligibility of clear and conversational speech. The results of such studies that have investigated differences

between conversational and clear speech are summarized below.

2.1.1 Intelligibility Differences

In a series of studies, Picheny, Braida, and Durlach investigated the differences between clear and conversational speech. The first study[44, 45] tested five hearing-impaired listeners on sets of 50 nonsense sentences spoken by three male talkers in both conversational and clear speaking modes. The sentences were presented at three different sound levels using two different frequency-gain characteristics. Intelligibility, based on key-word scores, was found to be 17 percentage points higher on average for clear speech than for conversational speech. Moreover, this intelligibility difference was independent of talker, listener, presentation level, and frequency-gain characteristic to a first approximation. This intelligibility advantage of clear speech over conversational speech has been verified[55, 28] and extended to include normal hearing listeners in noise[55, 42].

In a related study, Chen[6] investigated the intelligibility of conversationally and clearly spoken consonant-vowel (CV) syllables. The CV's were formed from one of the six stop consonants (/p/, /t/, /k/, /b/, /d/, /g/) followed by one of the three point vowels (/i/, /a/, /u/). Each CV was spoken both clearly and conversationally by three male talkers and presented to three normal hearing listeners in noise. On average, the intelligibility of clear speech was 22 percentage points higher than conversational speech, based on a CV-syllable correct score.

More recently, Payton, Uchanski, and Braida[42] examined the effects of noise and reverberation on intelligibility. This study consisted of nonsense sentences spoken clearly and conversationally presented in various environments to ten normal hearing and two hearing-impaired listeners. The environments were combinations of three levels of reverberation and four levels of noise, although not every environment was presented to every listener. On average, clear speech was 20 points more intelligible than conversational speech for normal hearing listeners and 26 points more intelligible for hearing-impaired listeners. In addition, this advantage was found to depend only on the intelligibility score for conversational speech but not on listener or environment.

2.1.2 Acoustic Differences

After establishing the high intelligibility of clear speech, Picheny, Braida, and Durlach[44, 46] went on to study the acoustical differences between conversational and clear speaking modes. They analyzed the acoustical properties of 50 nonsense sentences spoken clearly and conversationally by three male talkers. Substantial intratalker differences were observed for articulation rate, number of pauses, and number of phonological modifications. The short-term spectra of both consonants and vowels and also the relative intensities of plosives and fricatives were found to differ between clear and conversational speech. Although this study identified many acoustical differences between clear and conversational speech, it did not attempt to determine which differences were responsible for the high intelligibility of clear speech.

An acoustical analysis of clear and conversational speech was also performed in Chen's[6] study of CV-syllable intelligibility. Measurements demonstrated that clearly spoken syllables exhibited significantly longer voice onset times for voiceless consonants. Also, the formant frequencies of vowels were found to cluster more tightly in clear speech (in F1-F2 space), suggesting that the formants more closely approximated their target values. Clear speech also exhibited a larger vowel triangle, larger consonant-to-vowel ratios (ratio of the RMS levels of the consonant and vowel segments), and longer formant-transition durations.

2.1.3 Speaking Rate Differences and Effects on Intelligibility

Perhaps the most striking difference between clear speech (elicited without training) and conversational speech lies in speaking rate. The speaking rate for clear speech is usually only half that of conversational speech[44, 46]. As a result, several studies have attempted to determine whether a reduced speaking rate is essential to highly intelligible speech. For example, Picheny, Durlach, and Braida[44, 47] conducted a probe experiment to investigate the effect of overall speaking rate on intelligibility. Using Malah's algorithm[35], one male talker's clear sentences (rate = 100 wpm) were uniformly time-compressed to a typical conversational speaking rate of 200 wpm, and

his conversational sentences were uniformly expanded to meet typical clear speaking rates of 100 wpm. After this time-scaling of the waveforms, the processed sentences were presented to five hearing-impaired listeners. In both cases, the processed speech was less intelligible than the unprocessed speech. In a later study[54, 55], non-uniform time-scaling based on the Griffin-Lim algorithm[17] was used to process the sentences in order to determine the contribution of segmental-level durational differences between clear and conversational speech. Unfortunately, both hearing-impaired listeners in quiet and normal hearing listeners in noise found the processed sentences to be less intelligible than the unprocessed sentences. Although neither time-scaling procedure produced clear speech that was both more intelligible and no slower than unprocessed conversational speech, non-uniform time-scaling was less harmful to intelligibility than uniform time-scaling. In both studies, the intelligibility of twice-processed materials (i.e. sped clear speech expanded to its original rate of 100 wpm and slowed conversational speech compressed to its original rate of 200 wpm) indicated that most of the decrease in intelligibility was not due to signal processing artifacts. A possible explanation for the reduced intelligibility of the slowed speech is that neither the uniform nor the non-uniform time-scaling is likely to have captured very many of the acoustical differences between clear and conversational speech other than speaking rate. For example, deletions of phones and pauses that are present in conversational speech would not be restored by expansion, and time-scaling of clear speech could alter formant transitions in a way that may not occur naturally in speech.

In addition to studies of the effects of time-scaling, several studies have examined the role of pauses in clear speech. More frequent and longer pauses, in conjunction with lengthened speech sounds, are responsible for the reduced speaking rate of clear speech[46]. Choi[7] investigated whether these pauses contribute to the higher intelligibility of clear speech. Her results indicate that adding pauses to conversational speech does not improve its intelligibility, and deleting pauses from clear speech does not decrease its intelligibility. This finding is supported by a related study by Uchanski[54], in which key words excised from clearly spoken sentences had approximately the same intelligibility as the same words in (nonsense) sentence

context.

Since artificial methods for altering rate have not yet produced clear speech at normal speaking rates, some more recent studies have attempted to elicit clear speech naturally. For example, Uchanski[54, 55] employed a professional "fast" talker to produce clear speech at a variety of rates. Intelligibility measurements suggested that the talker could not improve his intelligibility without slowing down.

A related study[28], however, shows that with training, talkers can produce a form of clear speech at nearly normal speaking rates. In order to improve the chances of obtaining natural clear speech at normal speaking rates in this study, much attention was given to selecting talkers for the experiment. Fifteen talkers with a minimum of two years speaking experience were screened for their ability to control both clarity and speaking rate. The talkers were familiarized with the characteristics of clear speech and were asked to mimic clear speaking styles that had been presented. Their attempt at clear speech as well as their conversational speech was then presented in noise (SNR = -4dB) to normal hearing listeners. From the intelligibility results and speaking rate measurements, five talkers were selected for further training based on their ability to manipulate clarity and/or speaking rate. The talkers were trained using an interactive procedure that provided feedback on both speaking rate and intelligibility. In this procedure, a metronome specified the speaking rate for the talkers, and four normal hearing listeners provided feedback on the intelligibility of their speech. The talker's speech was presented to each listener in turn and was distorted by multiplicative noise[49] in order to facilitate the elicitation of highly intelligible speech. Initially, the SNR was set to 0 dB, and it was decreased in steps of 0.2 dB until the listeners reported on average no more than one key word correctly from the talker's first utterance of the sentence. The talker was required to repeat a sentence with increased emphasis on articulation until more than half of the key words in the sentences were perceived correctly by a listener. No special training was provided for eliciting conversational speech. After training, clear and conversational speech were obtained from the talkers at three different relative speaking rates: "slow," "normal," and "quick." In order to measure the intelligibility of the six elicited speaking styles

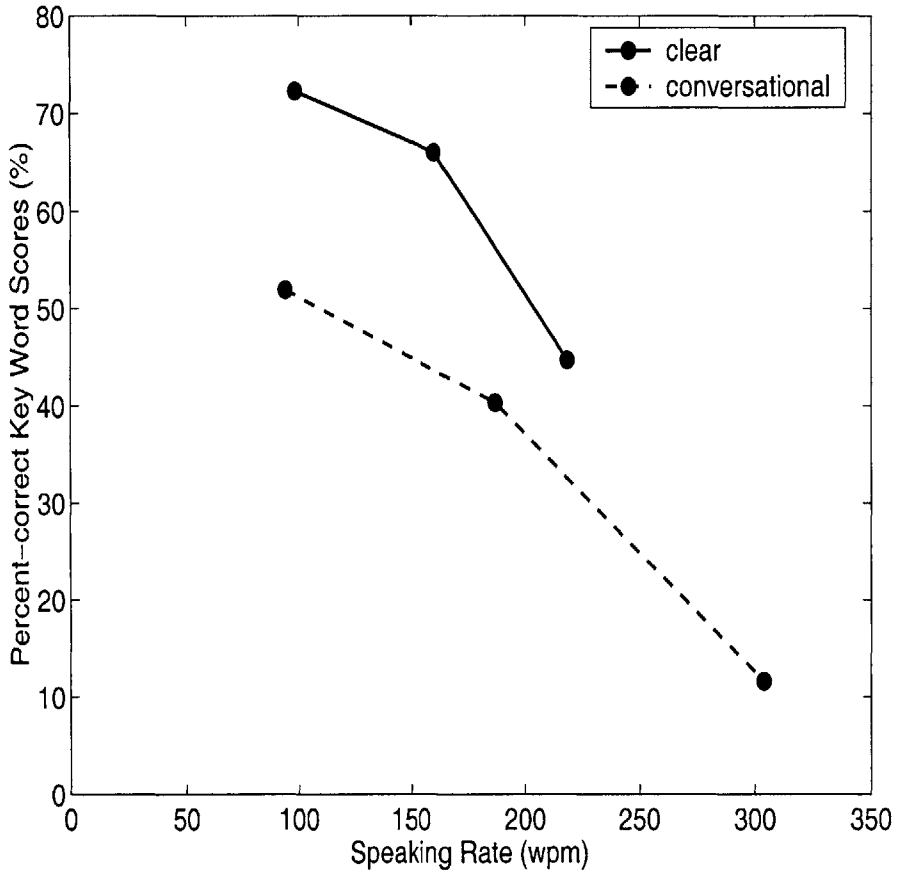

Figure 2-1: Average key-word scores for clear speech at slow, normal and quick speaking rate for all five talkers in Krause's[28] study.

(referred to by mode/rate as follows: clear/slow, conv/slow, clear/normal, conv/normal, clear/quick, conv/quick), the recordings were then presented to normal hearing listeners in the presence of additive speech-shaped noise[39], with the SNR set to -2dB. For all talkers, clear speech provided an intelligibility advantage over conversational speech in both clear/slow (18 percentage points on average) and clear/normal (14 points on average) speaking styles but not in the clear/quick style[28]. Moreover, the average speaking rate for clear/normal speech was 174 wpm, which is typical of conversational rates for these materials (178 wpm). This result, summarized in Figure 2-1, confirms that acoustical factors other than speaking rate can contribute to the high intelligibility of clear speech.

2.2 Previous Attempts at Intelligibility Enhancement

Since acoustical factors other than speaking rate contribute to the intelligibility advantage of clear speech over conversational speech, signal processing schemes that transform the conversational speech signal appropriately could improve intelligibility. A substantial portion of the previous work aimed at intelligibility enhancement has concentrated on techniques for processing degraded speech such as noise suppression (e.g., [27]) or dereverberation (e.g., [34]), rather than altering the undegraded speech signal itself. While these methods are valuable for improving speech intelligibility in some situations, most of the algorithms are designed for specific types of degradations and provide little benefit for other degradations. In contrast, processing the speech signal before degradations are introduced so that its acoustic properties are similar to that of clear speech could improve intelligibility for several types of degradation, because clear speech has been shown to provide an intelligibility advantage in several different conditions [45, 55, 42, 28, 29].

Some previous work in the area of intelligibility enhancement has focused on improving intelligibility by processing the undegraded speech signal. In an effort to improve radio communication technology, for example, a number of studies have evaluated processing techniques for improving the intelligibility of speech presented to listeners in high noise environments. Early studies[32, 30] showed that peak clipping the undegraded speech signal significantly improved intelligibility in high levels of additive noise. A subsequent study[38] demonstrated that high-pass filtering followed by rapid amplitude compression was a more effective processing scheme and improved speech intelligibility to the point where listeners could withstand a roughly 15 dB greater noise level than used for unmodified speech and still achieve intelligibility scores of 80 percentage points. This improvement in intelligibility is quite dramatic, but the intelligibility tests were performed only on normal hearing listeners.

While research in the area of manipulating the acoustic parameters of the speech signal is far from exhaustive, a number of studies have focused on processing specific

characteristics of the speech signal in an effort to improve intelligibility. One such study [4] implemented a signal processing method for altering spectral contrast, or the difference in amplitude of spectral peaks and valleys. In this study, short nonsense utterances were processed to three different levels of spectral contrast. Hearing-impaired listeners were asked to identify /b/, /d/, or /g/ at a fixed location in the utterance. Results for /b/ and /g/ showed a modest intelligibility gain as spectral enhancement increased, but results for /d/ were not consistent with this trend. In a study by Tallal *et al.*[51], a similar acoustical modification of speech was linked to significant improvements in speech discrimination and language comprehension abilities of language-learning impaired children. The acoustical modification, designed to make rapidly changing speech elements such as formant transitions more salient, consisted of lengthening the speech signal by 50% and differentially enhancing the amplitude envelopes in the 3-30Hz range by as much as 20 dB.

Other studies have investigated the effects on intelligibility of acoustic modifications of the speech signal to increase consonant-vowel ratio (CVR), or ratio of consonant power to the power in the nearest adjacent vowel. In particular, two studies reported modest intelligibility improvements associated with an artificial CVR increase for words[37] and nonsense syllables[15] presented to impaired listeners. A limitation of both studies is the lack of comparison to a linear amplification system with a frequency-gain characteristic appropriate for the impaired listener, which would tend to provide high frequency pre-emphasis and consequently increase CVR. These studies applied different amounts of gain to the consonant; one study applied a fixed gain and the other applied a gain sufficient to achieve a fixed CVR. Neither of these strategies represents the changes in CVR between clear/slow and conversational speech that have been reported[44]. Modifying consonant amplitudes to achieve CVR increases similar to those found in clear/slow and clear/normal speech could produce more significant gains in intelligibility, since clear speech is 17% more intelligible on average that conversational speech.

Before such modifications to enhance speech intelligibility were explored, however, the changes in acoustical characteristics between clear/normal and conv/nor-

mal speech were examined in order to determine the differences that remain between clear and conversational speech spoken at the same rate. The acoustical differences observed between conv/normal and clear/normal speech are discussed in Chapters 3 and 4, and subsequent speech processing attempts to enhance intelligibility based on these measured differences are described in Chapter 5. Theoretical and experimental measures of intelligibility used for evaluating the processed speech are presented in Chapter 6. Finally, a discussion of results and suggestions for future work is included in Chapters 7 and 8.

Chapter 3

Acoustical Analysis

To isolate the characteristics that contribute to highly intelligible, normal rate speech, an extensive acoustical analysis of both conversational and clear waveforms was performed. Although measurements of the acoustic differences between clear and conversational speech have been made previously by Picheny *et al.*[45] and Uchanski *et al.*[55], these studies were limited to clear speech that was produced at slower speaking rates than conversational speech. Thus, many of the acoustic differences (e.g. more frequent and longer pauses) reported by these investigators are likely to be related to differences in rate. A comparison of the acoustical properties of conv/normal and clear/normal speech could help identify factors responsible for the intelligibility advantage of clear speech, since these speaking styles differ only in speaking mode and not in speaking rate. In order to establish the relevant factors, a series of measurements was made that paralleled those of Picheny *et al.* and Uchanski *et al.*. To characterize acoustic differences between the styles of speech, acoustical measurements were examined at three levels of detail: global (sentence-level), phonological, and phonetic.

3.1 Speech Materials

All of the speech stimuli from Krause's study[28] were available for analysis. In that study, each set of 50 sentences was recorded in two speaking modes for each of five

talkers. The materials that pertain to clear and conversational speech for each talker consist of 300 utterances: one set of 50 sentences was recorded in conv/normal and clear/normal modes, one set was recorded in conv/normal and clear/slow modes, and one set was recorded in conv/slow and clear/slow modes. Although these stimuli are available for all five talkers in the study, materials from two of the talkers (RG and SA[28]) were selected for purposes of acoustical analysis based on their ability to vary both rate and clarity with more accuracy than the other talkers.

To facilitate acoustical measurements at the phonetic level, all 600 utterances (2 talkers x 3 sentences sets x 50 sentences/set x 2 utterance modes/sentence) were phonetically labeled (a pronunciation key for the labels used can be found in Appendix G). This was accomplished by combining information from aural repetitions and spectrograms. The phonetic labels for each of the utterances was also aligned with a dictionary pronunciation of each sentence so that phonemes could be compared across speaking mode and talker.

A subset of the phonetically labeled stimuli, consisting of 200 utterances, served as the primary database for the acoustical analysis. These 200 utterances were the 50 sentences that each talker recorded in both conv/normal and clear/normal modes, thus allowing direct phonetic comparisons between the two speaking modes. All acoustical measurements described below were conducted on the primary database. In addition, any measurements that could be automated (e.g., pause distribution) were also conducted on the 200 utterances recorded by the talkers in both conv/normal and clear/slow modes in order to facilitate comparisons with the findings reported by Picheny *et al.*[46].

3.2 Global Measurements

At the sentence level, differences in pause structure, fundamental frequency distribution, and long-term RMS spectra were measured. In addition, temporal envelope modulations were examined, since recent studies[50, 12, 11, 10] suggest that these modulations play an important role in speech intelligibility.

3.2.1 Pause Length Distribution

Conv/normal and clear/slow speech has been shown by Picheny et al.[46] to exhibit different patterns in pause length distribution. Specifically, clear/slow speech was found to have more frequent and longer pauses than conv/normal speech. However, since words excised from clear/slow sentences and presented to listeners in isolation do retain their high intelligibility[55], it does not seem likely that pause distribution plays a significant role in speech intelligibility. To verify this conclusion, pause frequency and duration were measured in the present study for both talkers in conv/normal, clear/normal, and clear/slow speaking modes. As in Picheny et al.[46], pauses were defined to be any period of silence of at least 10ms in duration, excluding silent intervals due to stop consonant closures.

The resulting pause length distributions are displayed in Figure 3-1. Appendix A also contains a summary of means and standard deviations for each talker and speaking style (see Table A.1). As in the previous work, a dramatic increase in pause frequency and duration for clear/slow speech was measured. However, the pause distribution for clear/normal speech was nearly the same as that of conv/normal speech. Thus, when constraints are placed on the talker's rate, the extraneous pauses found in clear/slow speech are eliminated. Consequently, it appears that increases in pause duration or frequency are not necessary components of highly intelligible speech.

3.2.2 Fundamental Frequency Distribution

Using the pitch estimation program provided in the *ESPS/waves+* software package, fundamental frequency (F0) values were extracted at 10ms intervals from voiced portions of the speech signal. A histogram of the F0 values obtained for both talkers in each speaking style is shown in Figure 3-2. Appendix A also contains a summary of means and standard deviations for each talker and speaking style (see Table A.2). SA exhibits a wider range in fundamental frequency as well as a somewhat higher average value in clear speech (at both normal and slow rates) than in conversational speech. Moreover, SA's F0 behavior for clear/slow speech is consistent with that previously

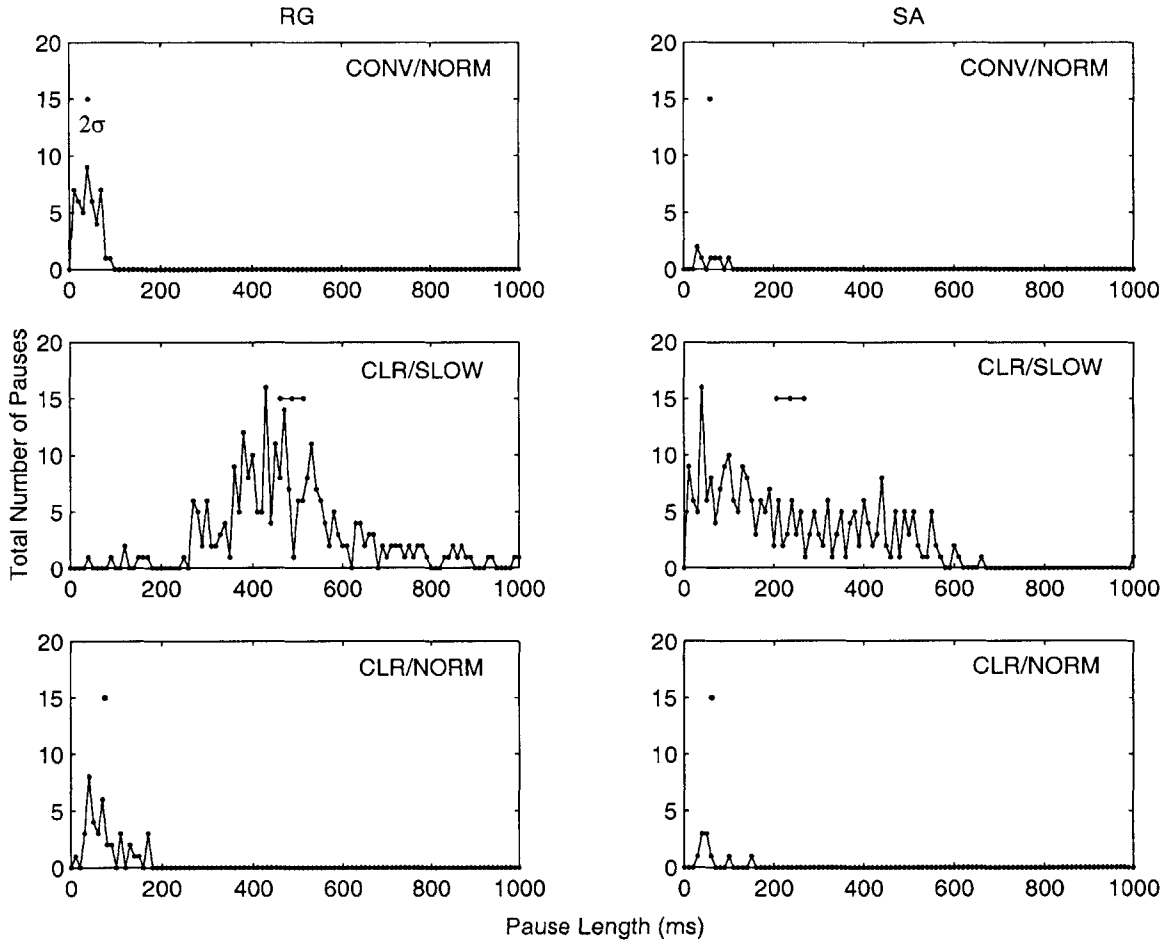

Figure 3-1: Pause length distributions. Each row shows distributions for different speaking modes; columns give results for each talker.

reported by Picheny et al.[46]. Although RG does not show a similar change in F0 average or range across speaking styles, her range of roughly 200Hz in F0 is consistent across all three speaking styles and quite large compared to those reported by Picheny et al.[46]. Gender differences may account for the differences in F0 behavior between talkers, since RG is female and SA is male. Females tend to have a wider range in fundamental frequency as compared to males. Bradlow *et al.*[3] found an average fundamental frequency range of 175Hz for a group of ten female talkers and 103Hz for a group of ten male talkers. Moreover, the female talkers had significantly higher intelligibility than the male talkers, as a group. Since none of the talkers in

the Picheny et al. study[45] were female, it was previously unknown whether a female talker must necessarily increase an already large F0 range when speaking clearly. Since the largest range of F0 that SA attained ($\approx$ 200Hz) was of roughly the same size as RG's range in all three speaking modes, and the maximum F0 range for any talker in the Bradlow *et al.*[3] study was 227Hz (exhibited by a female talker), it is possible that 200–225Hz may represent an upper limit on the amount that F0 can be varied without making the speech sound unnatural.

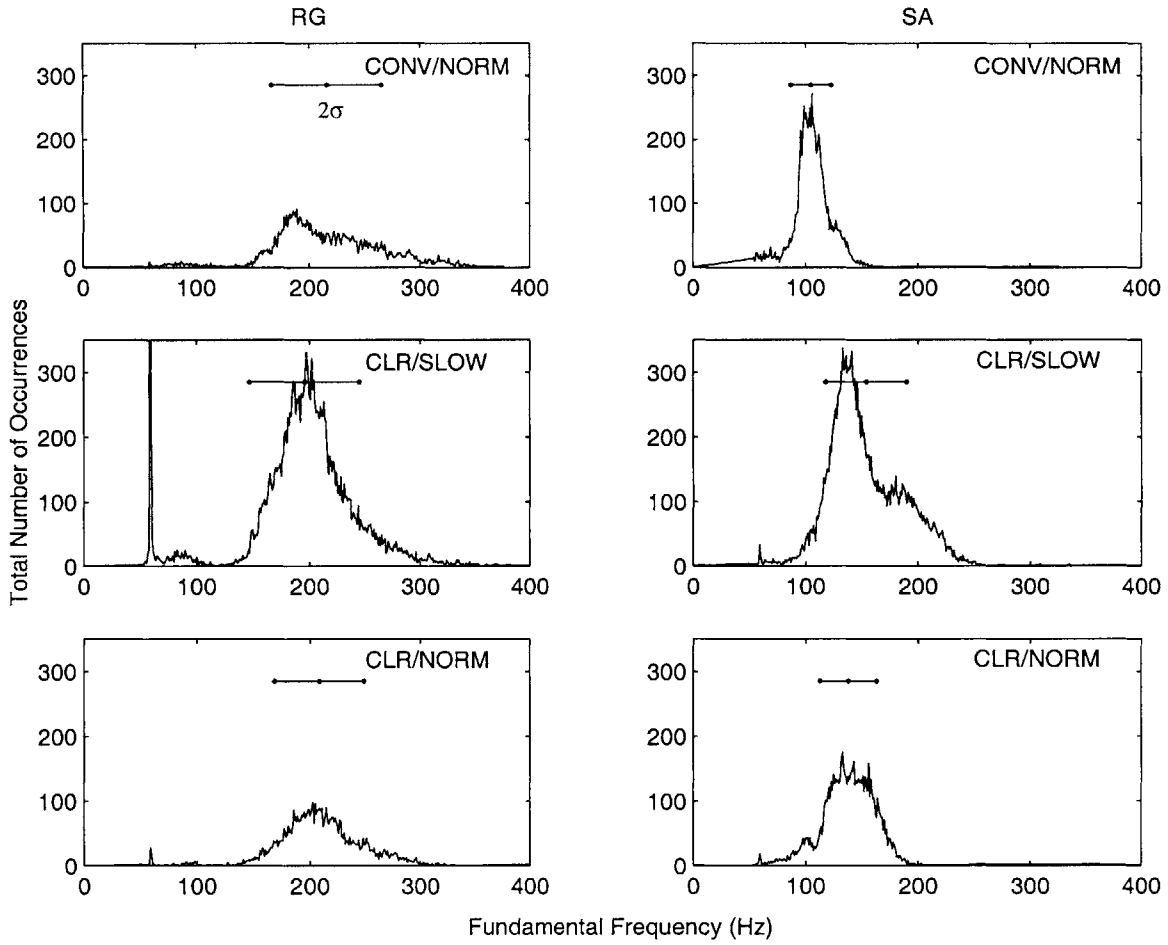

Figure 3-2: Fundamental frequency distributions. Each row shows distributions for different speaking modes; columns give results for each talker.

In addition to producing histograms of F0 values, a crude estimate of sentence level intonation contours was calculated by measuring F0 at two points in each sentence:

at its maximum value and 50ms before the end of the sentence. These values were averaged over 50 sentences for each speaking style. The results for both talkers are plotted in Figure 3-3. For SA, the maximum value is larger for clear/normal speech relative to conv/normal speech, and the value for clear/slow speech is even larger. However, there is little change in the sentence-final pitch across speaking styles. The results for SA are again similar to those reported by Picheny et al.[46]. The rise in SA's maximum values for clear/normal and clear/slow speech is likely due to the overall increase in pitch present in these styles. Since RG does not exhibit such an overall increase in pitch when speaking clearly, there is no corresponding increase in the maximum pitch value of the intonation contour.

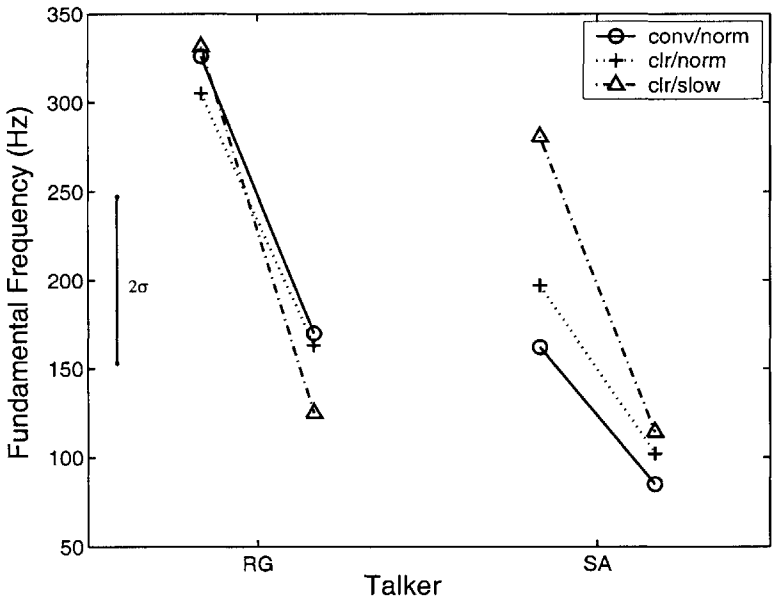

Figure 3-3: The maximum F0 value attained in a sentence followed by the value 50ms before the end of the last word in the same sentence, averaged over 50 sentences (a rough approximation of the F0 contour) for each speaking style.

3.2.3 Long-term Spectra

The long-term spectra of conv/normal, clear/normal, and clear/slow speech, normalized for long-term RMS level, were computed. For each speaking style, spectra were

averaged over 50 sentences. Using 25.6ms non-overlapping Hamming windows, FFTs were computed for each windowed segment within a sentence, and then the RMS average magnitude was determined. A 1/3-octave representation of the spectra was obtained by summing components over 1/3-octave intervals with center frequencies ranging from 62.5Hz to 8000 Hz. Finally, in order to examine the distribution of spectral energy for the clear modes relative to conv/normal speech, the conv/normal spectrum was subtracted from the clear/normal and clear/slow spectra. These spectral differences are shown in Figures 3-4 and 3-5, and the absolute spectra from which they were derived are presented in Figures A-1 and A-2 in Appendix A. For both talkers, the long-term spectrum of both clear/normal and clear/slow speech has relatively more energy above 1 kHz than conversational speech. A similar effect for clear/slow speech was reported by Picheny[46] for only one of three talkers and was not considered to be "substantial." However, the result is typical of a more pressed voice due to increases in vocal effort and vocal level, which are exhibited in the production of clear speech[46]. While an increase in vocal level has been shown to increase middle- and high-frequency components of the spectrum relative to low-frequency components[33], it has not been shown to produce large improvements in intelligibility[48]. Consequently, the high-frequency emphasis alone cannot explain the large intelligibility advantage of clear/normal speech over conversational speech. It may, however, be a contributing factor, since information contained in the high-frequencies is important for cueing place of articulation[36].

3.2.4 Temporal Envelope Modulations

Since the high-frequency emphasis found in clear/normal speech may play a role in cueing place of articulation for some consonants, the question arose whether a comparable emphasis on manner of articulation and voicing would exist. Because temporal envelope modulations are thought to be important for cueing manner and voicing[50], envelope modulations in clear/normal and conversational speech were investigated. In particular, the spectra of the octave band envelopes was examined.

To compute the envelope spectra, all 50 sentences in each condition were first

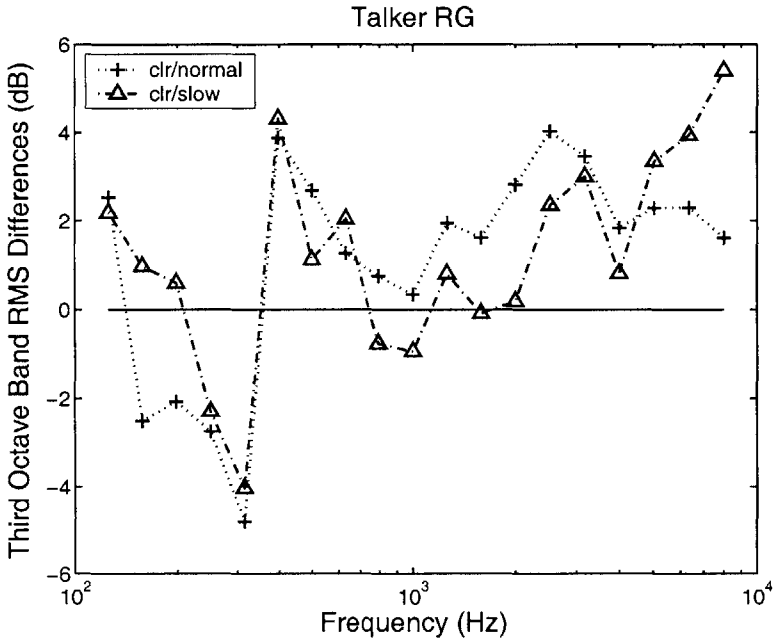

Figure 3-4: Third-octave band RMS spectral differences for RG, obtained by subtracting the conv/normal spectrum from the clear/normal and clear/slow spectra, depicts the relative distribution of spectral energy between conversational and clear speech.

concatenated to mimic running speech. The speech was then filtered into seven component signals, using a bank of 4th order octave-bandwidth Butterworth filters, with center frequencies 125Hz – 8000Hz. Since the sentences were digitized at a 20kHz sampling rate, the 8000Hz filter was implemented as a high-pass filter. The filter bank outputs for each of the seven octave-bands were then squared and low-pass filtered by an 8th order Butterworth filter with a 60Hz cutoff frequency in order to obtain relatively smooth intensity envelopes. Finally, the intensity envelopes were downsampled by a factor of 100, and power spectra were computed. A 1/3-octave representation of the spectra was obtained by summing components over 1/3-octave intervals with center frequencies ranging from 0.4Hz to 20 Hz. As in Houtgast and Steeneken[23], the power spectra were normalized by the mean of the envelope function. For a single 100% modulated sine-wave, this method would result in a value of 1.0 for the 1/3-octave band corresponding to the modulation frequency and zero for

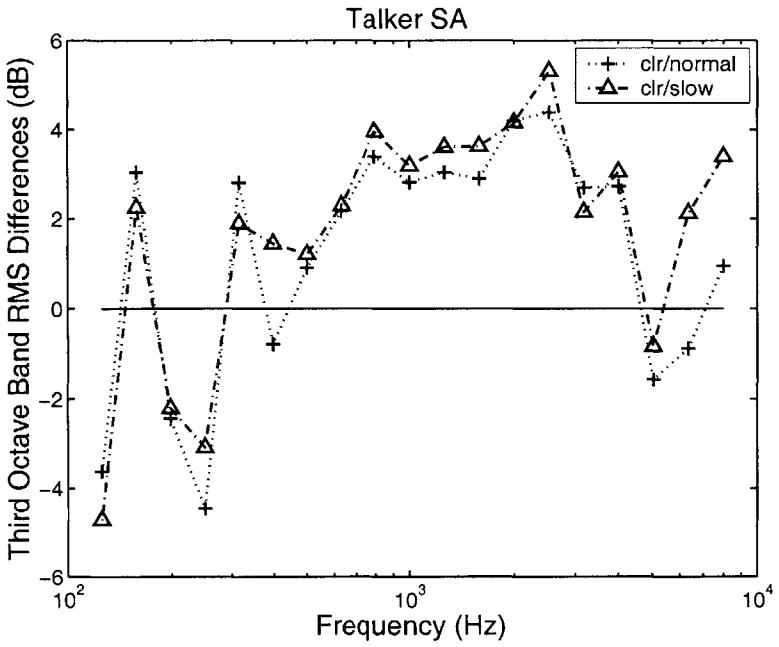

Figure 3-5: Third-octave band RMS spectral differences for SA, obtained by subtracting the conv/normal spectrum from the clear/normal and clear/slow spectra, depicts the relative distribution of spectral energy between conversational and clear speech.

the other bands. In other words, to the extent that only one modulation exists per 1/3-octave band, each normalized value can be considered the modulation index, m, for that band. The modulation index measures depth of modulation and is defined as the ratio K/A, where K denotes the maximum amplitude of the sinusoidal component of a signal, and A represents its DC component.

The spectra of the octave-band intensity envelopes for each talker are shown in Figures 3-6 and 3-7. In all of the octave-bands for both talkers, clear/slow speech has a higher modulation index than conv/normal speech for modulation frequencies up to 3–4Hz. A similar finding was reported previously by Payton et al.[42], but it was unknown whether the enhancement of slowly varying modulations was purely a result of the reduction in speaking rate associated with clear/slow speech. SA's data, however, shows a similar effect for clear/normal speech relative to conv/normal speech. For the 250Hz, 500Hz, 1000Hz, and 2000Hz bands, SA's clear/normal speech has a higher modulation index than his conv/normal speech over the same

range of modulation frequencies (< 4Hz). While the increase in modulation depth for clear/normal speech is not as great as in clear/slow speech and is present in only four of the seven octave bands, SA's strategy for producing clear speech at normal rates affects the intensity envelope spectra in a manner similar to his strategy at slow rates. The same is not true of RG, however. Only her clear/slow speech exhibits this trend.

In order to analyze the change in SA's envelope spectra for clear/normal speech from a time domain perspective, intensity envelopes of individual sentences were also examined. Figures 3-8 and 3-9, show typical intensity envelopes for conv/normal and clear/normal speech in the 500Hz and 1000Hz bands, respectively. In clear/normal speech, the envelopes of content words (nouns, verbs, and adjectives) were often of greater intensity relative to the envelopes of function words (words with primarily grammatical purpose). For example, in Figure 3-8 the content word "gold" had a higher overall intensity in clear/normal speech than in conv/normal speech, and while the other content words were roughly equal in intensity across the two speaking styles, the function words in clear/normal speech were reduced in intensity compared to their counterparts in conv/normal speech. As a result, the relative intensity of content words compared to function words was increased in clear/normal speech. This increase in the intensity of content words in clear/normal speech was particularly evident near the ends of sentences, as shown in Figure 3-9 where the envelope of the word "bluff" was of much higher intensity in clear/normal speech than in conv/normal speech.

Recent work by Drullman *et al.* suggest that a change in envelope spectra, such as the change measured between conv/normal and clear/normal speech for SA, may contribute to increased intelligibility. These studies[12, 11] demonstrated that modulations as low as 2Hz are most important for phoneme identification in CVC and VCV syllables. Although a slightly higher range (4–16Hz) was reported to contribute most to sentence intelligibility, the sentences used in the studies were contextual. Since little context is available when identifying words in nonsense sentences, the syllable data, rather than the sentence data, may be more relevant to the present work.

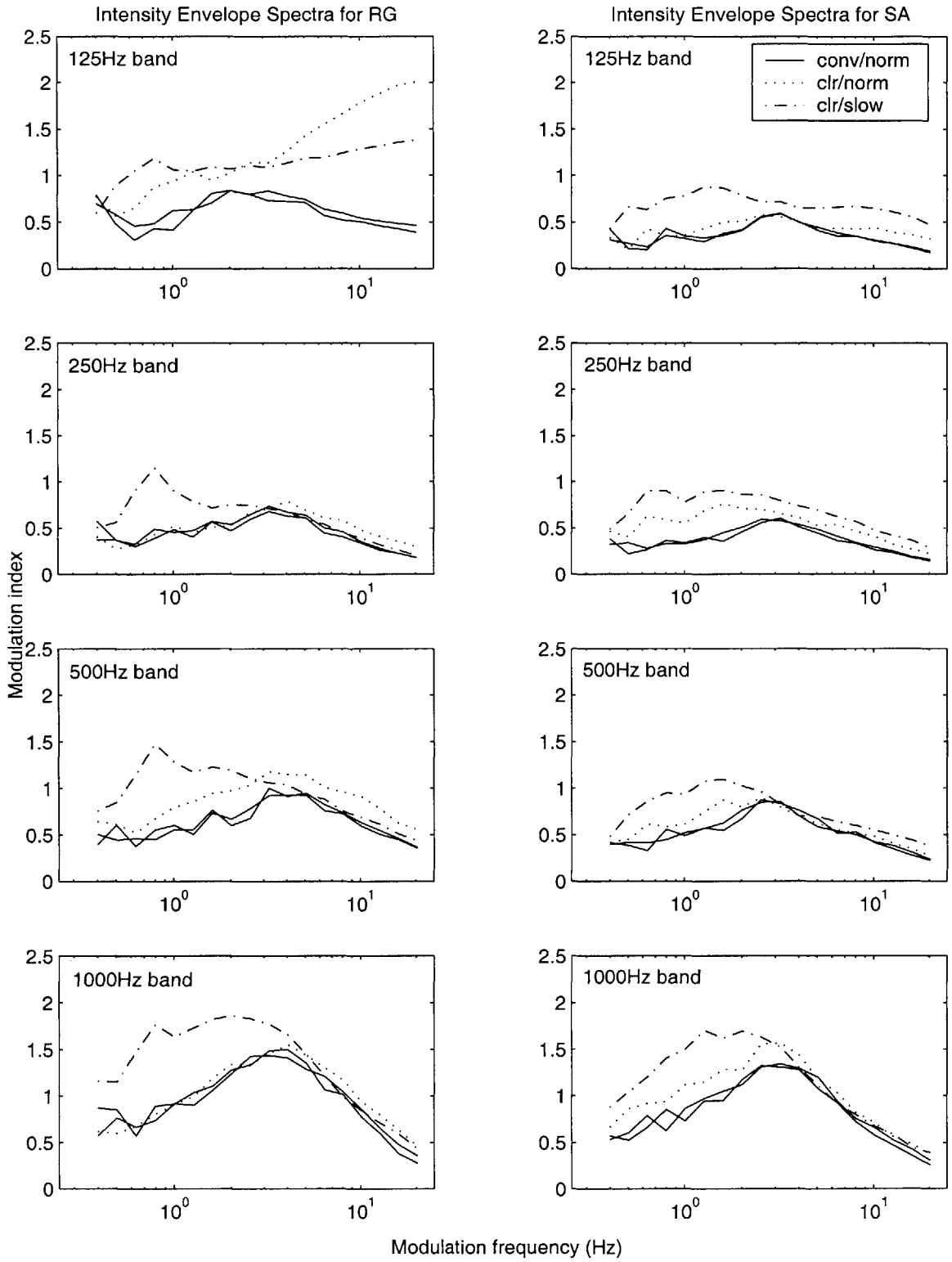

Figure 3-6: Spectra of intensity envelopes depicted by modulation index, indicating depth of modulation, as a function of third-octave band modulation frequency for Talkers RG and SA in lower four octave bands.

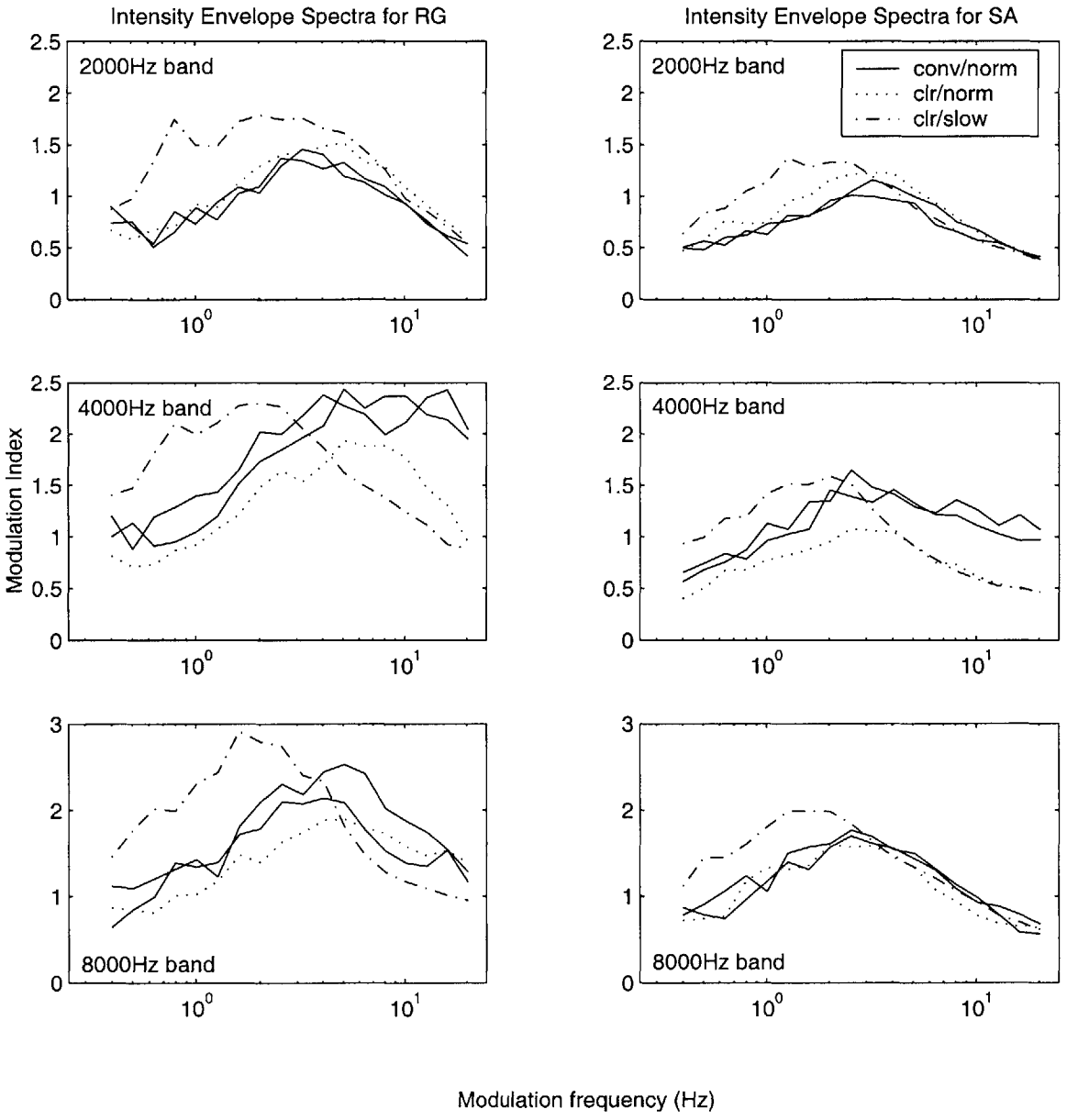

Figure 3-7: Spectra of intensity envelopes depicted by modulation index, indicating depth of modulation, as a function of third-octave band modulation frequency for Talkers RG and SA in upper three octave bands.

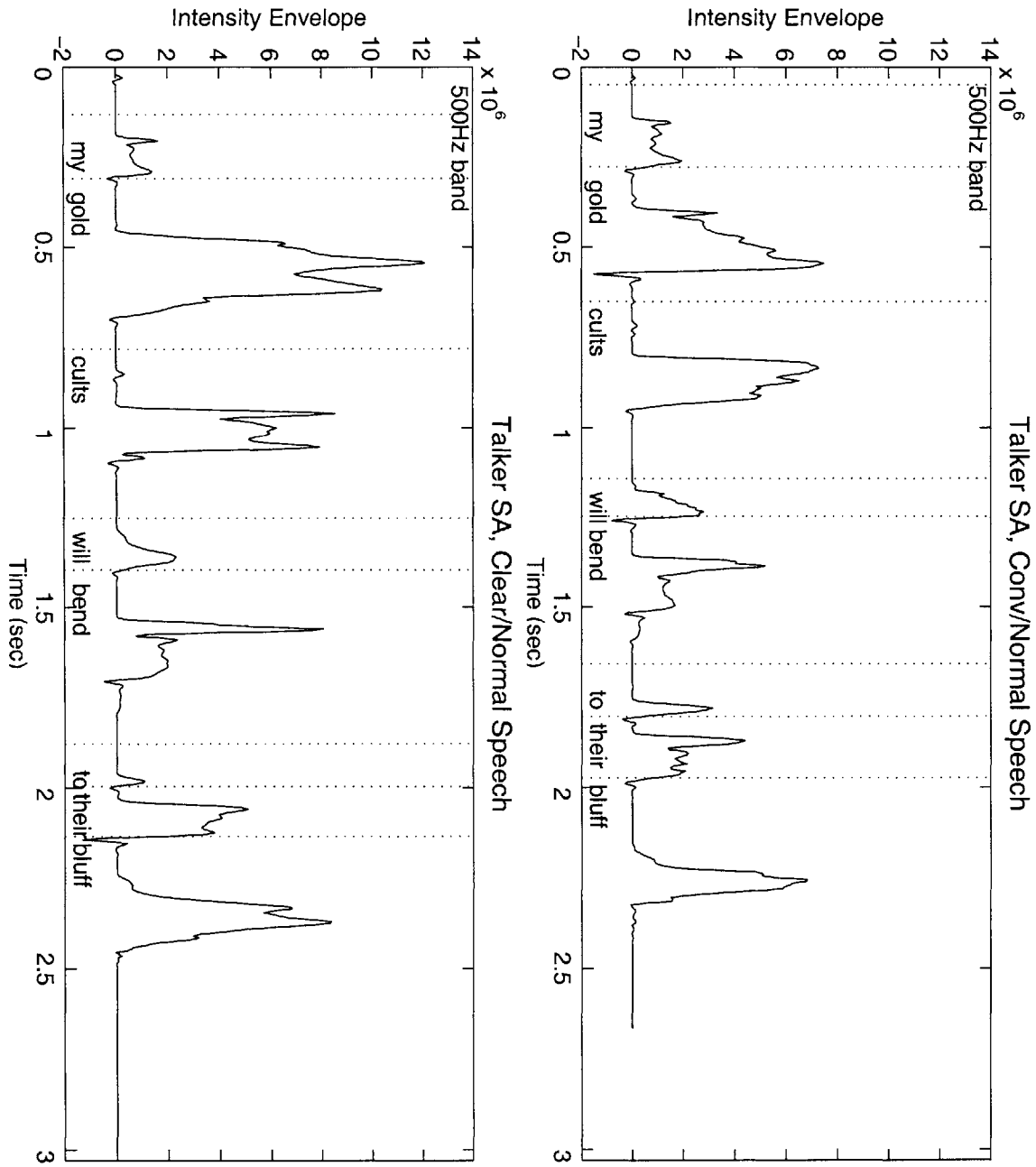

Figure 3-8: Intensity envelopes of SA's conv/normal and clear/normal speech in the 500Hz octave band for the sentence, "My gold cults will bend to their bluff." As in this sentence, envelopes of content words in clear/normal speech were typically more intense relative to envelopes of function words.

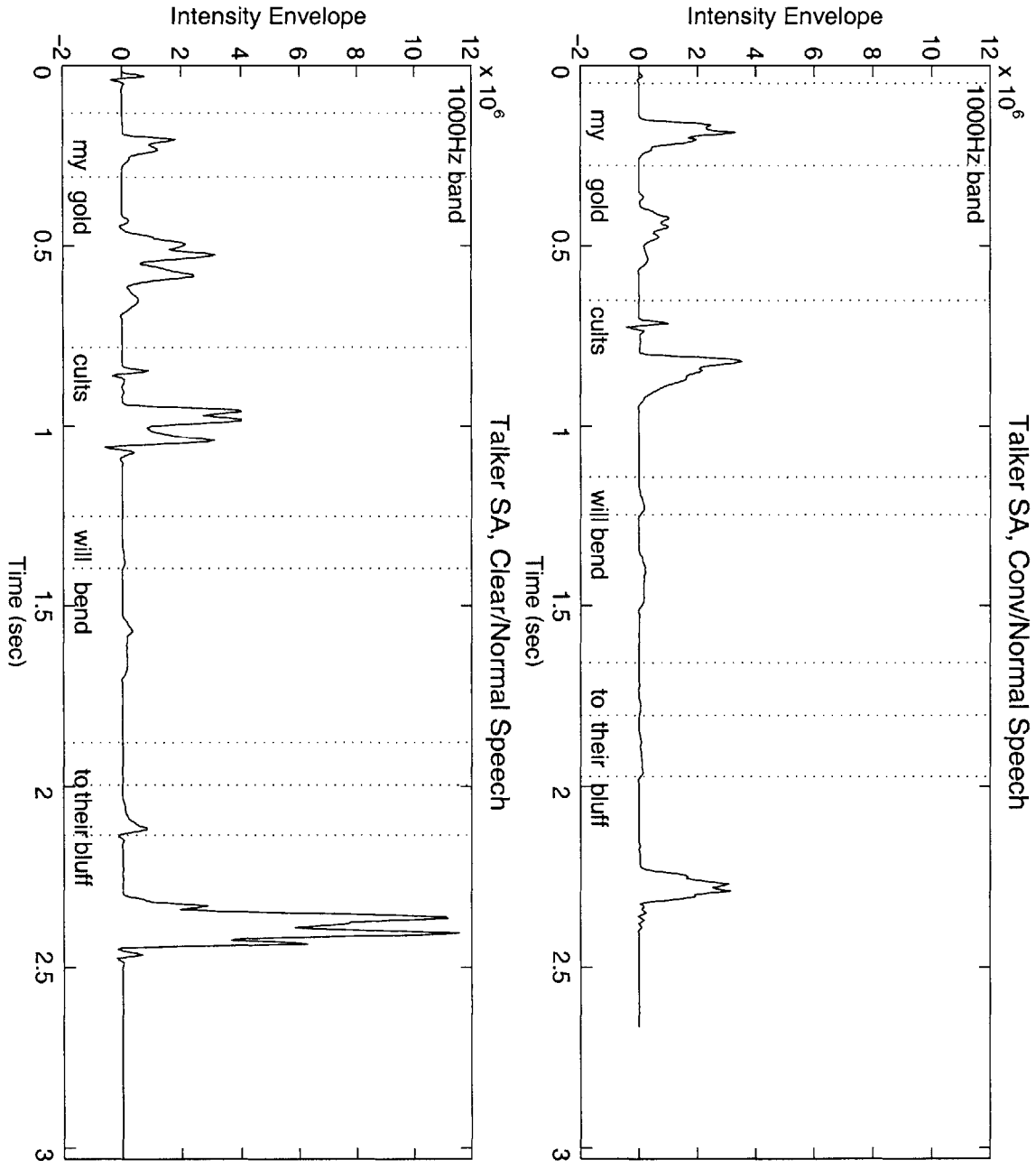

Figure 3-9: Intensity envelopes of SA's conv/normal and clear/normal speech in the 1000Hz octave band for the sentence, "My gold cults will bend to their bluff." As in this sentence, envelopes of content words in clear/normal speech, especially near the ends of sentences, were more intense relative to function words.

A second indication that the increase in lower modulation frequencies may contribute to improved intelligibility stems from the speech transmission index (STI). The STI is a measure of change in modulations of the intensity envelope due to degraded listening conditions. It has been shown to be highly correlated with speech intelligibility[23, 25, 42]. Although the STI is traditionally calculated from the measured change in modulation depth resulting from the transmission of modulated noise signals through an acoustic environment, methods for determining the STI directly from speech waveforms have also been proposed[22, 23, 41]. One such method was employed to determine the speech-based STI for the conv/normal, clear/normal, and clear/slow stimuli used in the original intelligibility experiment[28]. In this case, it was necessary to calculate intensity envelope spectra for the noisy stimuli (SNR $= -4dB$) presented to listeners in the intelligibility tests as well as the undegraded speech waveforms (see Figures 3-6 and 3-7). The modulation transfer function (MTF) was then obtained by taking the ratio of the degraded envelope spectra to the undegraded envelope spectra for each of the seven octave bands. From this speech-based MTF, the STI was calculated as in Houtgast and Steeneken[23].

The results of the STI calculations are displayed in Figure 3-10. The speech-based STI clearly differentiates the relative intelligibility of talkers and modes. Since an STI range of 0 to 1 should correspond with intelligibility scores ranging from 0% to 100%, the range of STI scores found here (0.07) is compressed relative to the range of intelligibility measures (30%). Nonetheless, the STI is highly correlated with the measured intelligibility scores, with a correlation coefficient of 0.90. Because the difference in intensity modulations between talkers and modes is so highly correlated with measured intelligibility, it is well worth investigating whether artificially manipulating the modulations will result in enhanced intelligibility. This question is pursued in Chapter 5.

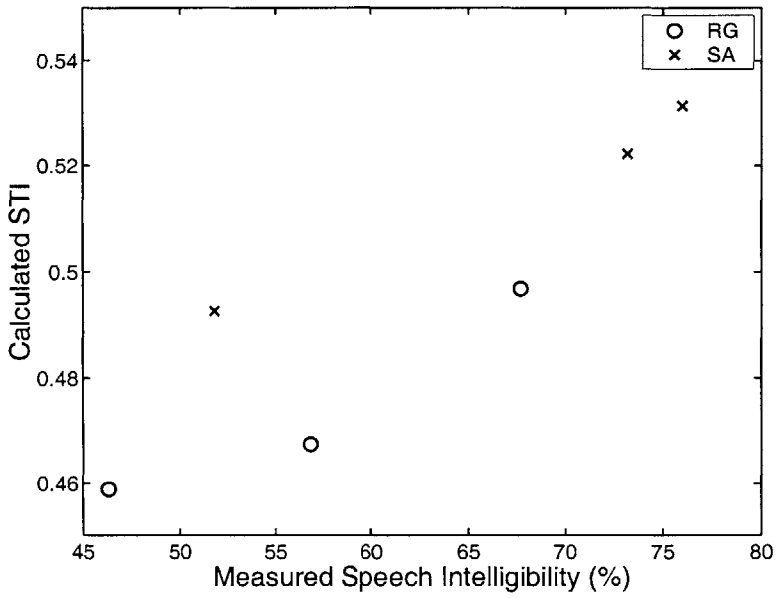

Figure 3-10: Measured intelligibility vs. calculated STI (for normal-hearing listeners) for both talkers in conv/normal, clear/normal, and clear/slow speaking modes.

3.3 Phonological Measurements

In running speech, speech segments in certain contexts undergo specified transformations described by phonological rules. The frequency of such phenomena occurring in clear and conversational speech was measured as a percent of total possible occurrences. Similar to Picheny et al.[46], the phenomena were classified into the following categories:

1. **Vowel modification (VM):** This category includes vowel substitutions and the reduction of vowels to a schwa in unstressed syllables. Also included in this category is the replacement of a vowel-sonorant sequence with a syllabic version of the sonorant. Total possible occurrences included a count of vowels in function words, unstressed vowels in content words, and vowel-sonorant sequences.

2. **Burst elimination (BE):** Burst elimination is the failure to create a burst at the release of a stop consonant. It is particularly common when the stop consonant is in word-final position or syllable-final position. Total possible occurrences included sentence-final stop consonants and word-final stop consonants followed by another consonant.

3. **Alveolar flap (AF):** This phenomenon occurs most frequently when a /t/ or /d/ is preceded by a stressed vowel and followed by an unstressed vowel. The /t/ or /d/ is frequently realized as a flap. Total possible occurrences included /d/ or /t/ occurring between two vowels.

4. **Sound insertion (SI):** This category refers to the insertion of a schwa vowel after a voiced consonant. The insertions frequently occur at the end of words, particularly in clear/slow speech[46]. Total possible occurrences was set to be the total number of syllables in the list.

Figure 3-11 shows histograms of these types of phonological modifications for both talkers in all of the speaking styles. Every histogram compares the same sentence list spoken in two different styles, but each of the four histograms is measured on a different list. Based on these results, it is likely that the method for counting total possible occurrences is somewhat crude, since markedly different scores are obtained for each talker's conv/normal speech on the two different sentence lists. While the range of alveolar flap scores for conv/normal speech is likely a result of the small number of possible occurrences (ranging from 5 to 10 for the various lists), the range of burst elimination scores cannot be accounted for with this reasoning. The most likely explanation is that global factors such as word position within the sentence influences a talker's decision to eliminate the burst. If this is the case, comparisons of modifications between modes on the same sentence list would have value, but comparisons across sentence lists would not be advised.

Comparing within sentence list, the results for clear/slow speech are similar to those reported by Picheny et al.[46]: the frequency of vowel modifications, eliminations of stop bursts, and alveolar flaps is smaller in clear/slow speech than in

conv/normal speech, and the number of sound insertions is much larger. In contrast, clear/normal speech has about the same number of sound insertions as conv/normal speech. Thus, it would appear that the sound insertions found in clear/slow speech are not an essential component of highly intelligible speech.

Perhaps the most striking finding regarding phonological modifications is that, while the trend for RG involving vowel modification, burst elimination, and alveolar flaps in clear/normal speech is similar to that of clear/slow speech, SA's data actually exhibits the opposite effect. The number of vowel modifications and burst eliminations in SA's clear/normal speech is higher than in his conv/normal speech. One possible explanation for this result is that the occurrences of these phenomena may have been concentrated primarily in function words, which were not scored in the intelligibility tests. Another possible explanation of this difference between talkers, in conjunction with other differences described above, is that the talkers do not employ the same strategies for producing clear speech at normal rates. Instead, each may have retained a somewhat different subset of the acoustical characteristics present in clear/slow speech. SA retained differences in F0 distribution and temporal envelope modulations between conv/normal and clear/normal speech, while RG retained this reduction in phonological phenomena. The idea of differing talker strategies will be explored further in Chapter 4.

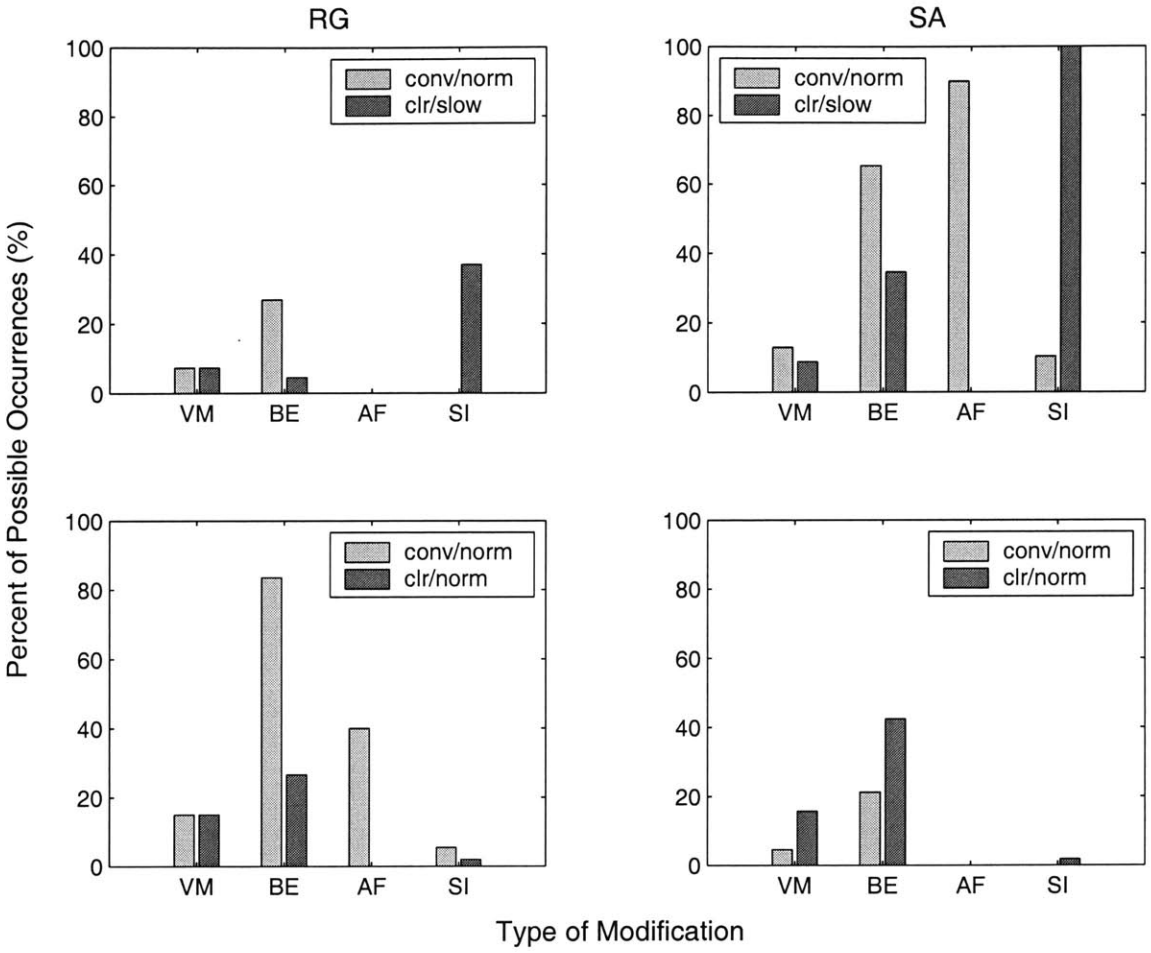

Figure 3-11: Frequency of phonological phenomena (**VM**=vowel modification, **BE**=burst elimination, **AF**=alveolar flap, **SI**=sound insertion). Top row compares sentences spoken in both conv/normal and clear/slow modes, and bottom row compares sentences spoken in both conv/normal and clear/normal modes. Each row is a different corpus of sentences. Columns give results for each talker.

3.4 Phonetic Measurements

Phonetic level measurements included segmental power, segmental phone duration, short-term RMS spectra, vowel formant frequencies, consonant-vowel ratio, and voice-onset time. Formant transition duration and extent was also examined, since some recent work[40, 51, 52] shows this parameter to be an important determinant of speech intelligibility. Unless otherwise stated, these measurements were made for each phonetic segment listed in Appendix G.

3.4.1 Power

To compare the relative power of phones spoken in conv/normal and clear/normal styles, the RMS level for each phonetic segment in both styles was computed. Segment boundaries necessary for the calculation were derived from the appropriate label files (see Section 3.1). Statistical comparisons of the conv/normal and clear/normal levels for each phone were performed via paired t-tests, where the difference variable $D = L_{conv} - L_{clear}$ was formed from L_{conv}, the level for the phone spoken conversationally, and L_{clear}, the level for the phone spoken clearly.

Phones with levels that differed significantly (p=0.05) are listed in Tables 3.1 and 3.2. Clearly, few phones for either talker showed a statistically significant difference in level, considering that RMS levels were measured for 43 phones. This result is not particularly surprising, since the long-term RMS level of each sentence was normalized. For a talker to produce a particular phone with greater power in clear/normal speech relative to conv/normal speech would mean that one or more of the other phones would have to be produced with relatively less power. Talkers may have found experimentally that modifying relative segmental power was not beneficial to intelligibility, or perhaps the mental load and articulatory demands of producing speech in this manner was prohibitively difficult. In either case, it appears that a change in relative power of phonetic segments is not integral to highly intelligible speech.

The average energy of phones in conv/normal and clear/normal speech was also

Table 3.1: Segmental power data (means in dB) for RG in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Phone	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	v	Sig Level
AA	69.3	68.1	1.2	1.7	0.4	0.731	0.000	3.34	22	0.003
EH	69.6	68.5	1.1	3.0	0.5	0.455	0.007	2.24	33	0.032
EY	66.5	67.3	-0.8	2.0	0.3	0.726	0.000	-2.30	33	0.028
OW	70.5	68.8	1.7	1.4	0.3	0.718	0.000	5.98	24	0.000
UH	67.0	68.6	-1.7	2.2	0.6	0.771	0.002	-2.76	12	0.017
S	58.6	57.0	1.6	4.1	0.5	0.226	0.091	2.94	56	0.005
TH	53.4	47.6	5.9	4.0	1.8	0.766	0.131	3.30	4	0.030
V	55.2	59.9	-4.7	3.0	1.3	0.253	0.681	-3.48	4	0.025
W	61.8	64.7	-2.9	4.2	0.9	0.320	0.137	-3.31	22	0.003
L	64.8	65.7	-0.9	2.7	0.4	0.590	0.000	-2.29	42	0.027
M	59.7	63.2	-3.6	3.0	0.5	0.735	0.000	-6.67	30	0.000
N	58.6	62.6	-4.0	2.8	0.4	0.669	0.000	-9.00	38	0.000

calculated from level and duration measurements, according to the formula $E = L^2 \times T$, where E is energy, L is RMS level of the phonetic segment, and T is the length of the phonetic segment in seconds. The relative energy in each speaking style was compared for each phone via paired t-tests, where the difference variable was defined as $D = E_{conv} - E_{clear}$, the difference between the energy of the phone spoken conversationally (E_{conv}) and the energy of the phone spoken clearly (E_{clear}). Phones that differed significantly (p=0.05) in energy are listed in Tables 3.3 and 3.4. As with level, few phones for either talker showed a statistically significant difference in energy.

3.4.2 Duration

The relative durations of phones spoken in conv/normal and clear/normal styles was also computed and compared statistically via paired t-tests. In this case, the difference variable was $D = D_{conv} - D_{clear}$, where D_{conv} was the duration of the phone spoken conversationally, and D_{clear} was the duration of the phone spoken clearly.

Table 3.2: Segmental power data (means in dB) for SA in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Phone	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	υ	Sig Level
IH	66.7	67.4	-0.7	2.1	0.3	0.700	0.000	-2.28	50	0.027
IY	66.1	67.8	-1.6	2.5	0.6	0.331	0.180	-2.74	17	0.014
Z	57.4	58.5	-1.1	2.4	0.3	0.484	0.001	-3.08	46	0.003
Y	62.7	65.2	-2.5	2.5	0.6	0.537	0.018	-4.38	18	0.000
W	63.3	65.0	-1.7	1.6	0.4	0.714	0.003	-3.93	14	0.002
R	67.5	68.2	-0.7	1.9	0.2	0.445	0.000	-3.31	77	0.001
M	63.0	65.7	-2.7	1.8	0.3	0.602	0.000	-9.79	41	0.000
N	61.9	63.6	-1.6	4.6	0.7	0.491	0.000	-2.43	46	0.019

Phones with durations that differed significantly (p=0.05) between speaking modes are listed in Tables 3.5 and 3.6. Again, not a great many phones of the 43 measured showed a statistically significant difference in duration for either talker. Moreover, although the phones that did change duration significantly for both talkers were primarily vowels, there was not a consistent pattern of change across talkers. Only three phones (/ah/, /eh/, and /z/) appeared on both talkers' lists, and the changes in duration for /ah/ and /eh/ were in opposite directions for each of the talkers. This trend was true for all vowels with a significant change in duration: RG's vowels were shortened in clear/normal speech relative to conv/normal speech, while SA's vowels were lengthened.

Table 3.3: Average segmental energy data (means in dB) for RG in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Phone	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	υ	Sig Level
EY	175.1	177.2	-2.1	4.2	0.7	0.719	0.000	-2.97	33	0.006
OW	184.0	181.5	2.5	3.3	0.7	0.816	0.000	3.78	24	0.001
UH	164.8	170.8	-6.0	5.1	1.4	0.828	0.000	-4.23	12	0.001
S	157.2	153.6	3.6	8.5	1.1	0.452	0.000	3.16	56	0.003
TH	141.1	133.6	7.5	5.9	2.7	0.912	0.031	2.83	4	0.047
W	159.7	166.6	-6.9	7.7	1.6	0.526	0.010	-4.27	22	0.000
R	168.5	171.0	-2.4	5.7	0.7	0.610	0.000	-3.60	71	0.001
L	167.2	170.1	-2.9	6.3	1.0	0.418	0.005	-2.97	42	0.005
M	156.3	162.7	-6.4	7.4	1.3	0.721	0.000	-4.83	30	0.000
N	152.7	161.4	-8.7	6.8	1.1	0.619	0.000	-8.01	38	0.000

Table 3.4: Average segmental energy data (means in dB) for SA in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Phone	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	υ	Sig Level
AH	174.1	170.1	4.0	11.2	1.9	0.499	0.002	2.13	34	0.040
UH	172.0	168.6	3.5	4.8	1.3	0.718	0.006	2.63	12	0.022
UW	180.9	177.6	3.3	4.1	1.2	0.876	0.000	2.79	11	0.018
Y	161.2	165.8	-4.6	6.8	1.6	0.463	0.046	-2.97	18	0.008
W	163.4	169.1	-5.7	4.5	1.2	0.732	0.002	-4.89	14	0.000
R	172.9	174.3	-1.4	4.5	0.5	0.592	0.000	-2.75	77	0.007
M	164.8	170.1	-5.3	3.2	0.5	0.798	0.000	-10.70	41	0.000
H	154.3	149.6	-4.8	11.6	2.1	0.770	0.000	2.29	30	0.030

Table 3.5: Segmental duration data (means in milliseconds) for RG in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Phone	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	υ	Sig Level
AA	90.2	105.7	-15.5	31.6	6.6	0.724	0.000	-2.35	22	0.028
AH	78.3	94.3	-16.1	27.9	6.6	0.752	0.000	-2.44	17	0.026
AO	110.8	133.0	-22.2	15.5	6.3	0.985	0.000	-3.51	5	0.017
AY	136.9	152.0	-15.1	22.2	4.7	0.960	0.000	-3.20	21	0.004
EH	89.4	102.6	-13.2	30.0	5.1	0.772	0.000	-2.56	33	0.015
IH	74.6	85.4	-10.8	24.4	3.6	0.689	0.000	-3.01	45	0.004
OW	148.5	162.8	-14.3	28.6	5.7	0.790	0.000	-2.50	24	0.020
UH	36.1	48.2	-12.2	8.8	2.4	0.504	0.079	-4.99	12	0.000
Z	95.4	78.4	17.1	27.6	4.7	0.846	0.000	3.60	33	0.001
R	66.3	75.3	-9.0	26.1	3.1	0.619	0.000	-2.93	71	0.004

Table 3.6: Segmental duration data (means in milliseconds) for SA in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Phone	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	υ	Sig Level
AH	90.6	83.5	7.1	18.7	3.2	0.910	0.000	2.24	34	0.032
EH	102.1	91.6	10.6	18.1	3.0	0.920	0.000	3.51	35	0.001
EY	205.1	169.2	35.9	55.4	10.3	0.876	0.000	3.49	28	0.002
ER	130.1	116.7	13.4	24.7	5.3	0.974	0.000	2.54	21	0.019
IY	156.1	130.4	25.6	38.1	9.0	0.823	0.000	2.85	17	0.011
OW	158.4	136.1	22.3	52.8	10.4	0.642	0.000	2.15	25	0.041
UW	207.7	149.3	58.4	37.1	10.7	0.946	0.000	5.45	11	0.000
DH	52.2	28.3	13.9	26.2	4.9	0.517	0.005	2.81	27	0.009
S	161.5	141.1	20.4	32.6	3.8	0.881	0.000	5.30	71	0.000
Z	118.4	105.6	12.8	23.9	3.5	0.924	0.000	3.66	46	0.001
W	95.4	78.4	-20.1	36.0	9.3	0.602	0.018	-2.17	14	0.048
L	95.4	78.4	11.7	40.1	5.7	0.645	0.000	2.05	48	0.046
N	95.4	78.4	10.7	26.2	3.8	0.882	0.000	2.78	46	0.008
JH	95.4	78.4	39.7	32.1	13.1	0.960	0.002	3.03	5	0.029
H	95.4	78.4	12.9	32.6	5.8	0.762	0.000	2.20	30	0.036

The above measurements, however, were made for all phones, independent of phonetic context. Since it is well known that vowels preceding voiceless consonants (in the same syllable) are typically shorter than vowels preceding voiced consonants[31, 8], durational measurements were also made taking context into account. Durations were measured for each vowel and for all vowels as a group, when followed by either voiced phones (/z/, /zh/, /v/, /dh/, /b/, /d/, and /g/) or unvoiced phones (/s/, /sh/, /f/, /th/, /p/, /t/, and /k/). Durations were compared statistically with paired t-tests as above. The results are shown in Tables 3.7 and 3.8. The trend is quite similar to the context-independent duration measurements, with RG showing a decrease in duration for clear/normal speech and SA showing an increase. This difference is most likely another indication of different talker strategies for speaking clearly at normal rates.

Table 3.7: Context-dependent duration data (means in milliseconds) for RG in conversational and clear modes at normal rate. "ALL" represents duration of all vowels as a group. Voiced consonants are indicated by "-V" and unvoiced consonants by "-U." Table shows only phones that were significant in paired t-tests at the p=0.1 level.

Phone Pair	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	v	Sig Level
ALL-V	102.3	118.1	-15.8	29.7	2.8	0.801	0.000	-5.67	112	0.000
AA-V	109.3	152.6	-43.3	14.8	5.6	0.874	0.010	-7.76	6	0.000
AH-V	73.8	89.6	-15.9	15.0	5.3	0.864	0.006	-3.00	7	0.020
AO-V	159.5	193.0	-33.5	3.5	2.5	1.000	0.000	-13.40	1	0.047
EH-V	108.5	128.1	-19.6	29.2	8.4	0.751	0.005	-2.32	11	0.040
ER-V	85.5	95.8	-10.4	22.5	5.5	0.797	0.000	-1.90	16	0.076
IH-V	67.7	90.1	-22.4	21.4	6.2	0.754	0.005	-3.63	11	0.004
ALL-U	95.9	110.0	-14.1	25.4	2.9	0.928	0.000	-4.79	73	0.000
AE-U	188.0	205.8	-17.8	18.2	8.2	0.874	0.052	-2.18	4	0.094
AY-U	130.5	165.5	-35.0	15.7	7.9	0.993	0.007	-4.45	3	0.021

Table 3.8: Context-dependent duration data (means in milliseconds) for RG in conversational and clear modes at normal rate. "ALL" represents duration of all vowels as a group. Voiced consonants are indicated by "-V" and unvoiced consonants by "-U." Table shows only phones that were significant in paired t-tests at the p=0.1 level.

Phone Pair	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	v	Sig Level
ALL-V	118.1	110.3	7.8	22.1	2.3	0.925	0.000	3.44	94	0.001
IY-V	119.0	107.8	11.3	8.2	4.1	0.991	0.009	2.74	3	0.072
UW-V	200.0	138.5	61.5	36.1	18.1	0.988	0.012	3.40	3	0.042
ALL-U	130.8	117.5	13.3	33.6	3.7	0.931	0.000	3.59	81	0.001
EH-U	172.0	155.5	16.5	11.4	5.7	0.984	0.016	2.88	3	0.063

3.4.3 Short-term RMS Spectra

The short-term spectra of conv/normal and clear/normal speech, normalized for segment RMS level, were computed. For both normal rate speaking modes, spectra corresponding to each phone were averaged over every occurrence of the phone in 50 sentences. The spectra were computed by averaging FFTs of windowed portions of the signal using 25.6ms Hamming windows, incremented in 1ms intervals. The spectra were also passed through a pre-emphasis filter with a slope of 6dB/octave in order to boost the higher frequencies. Finally, a 1/3-octave representation of the spectra was obtained by summing components over 1/3-octave intervals with center frequencies ranging from 62.5Hz to 8000 Hz. Typical results for consonants are shown in Figures 3-12 and 3-13. Although there is a slight difference between talkers, neither talker exhibits much spectral change between conv/normal and clear/normal styles. The results for vowels, however, are more interesting.

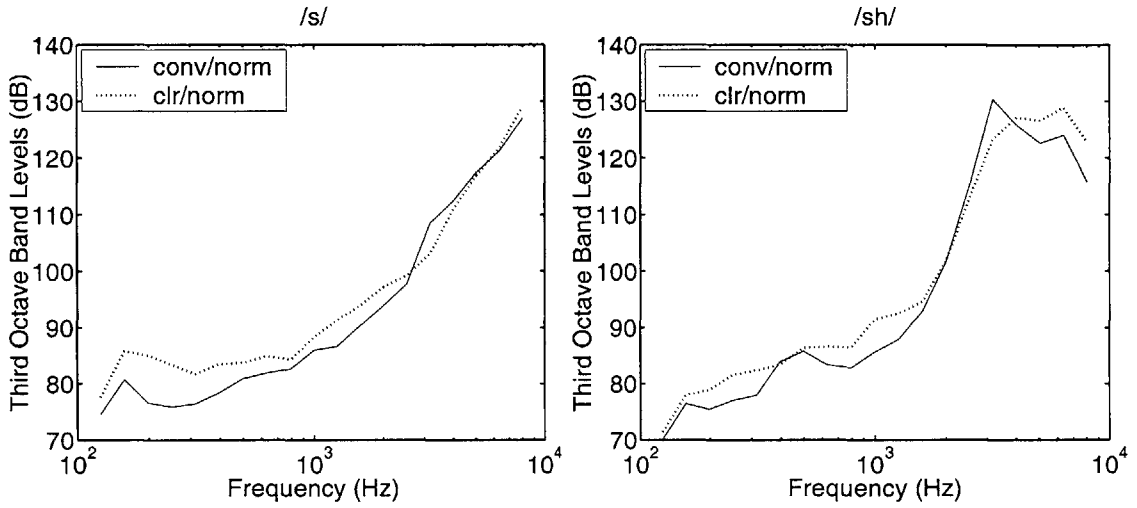

Figure 3-12: Third-octave average spectra of /s/ and /sh/ in conv/normal and clear/normal modes for RG. Similar results (no significant difference between modes) were obtained for all consonants.

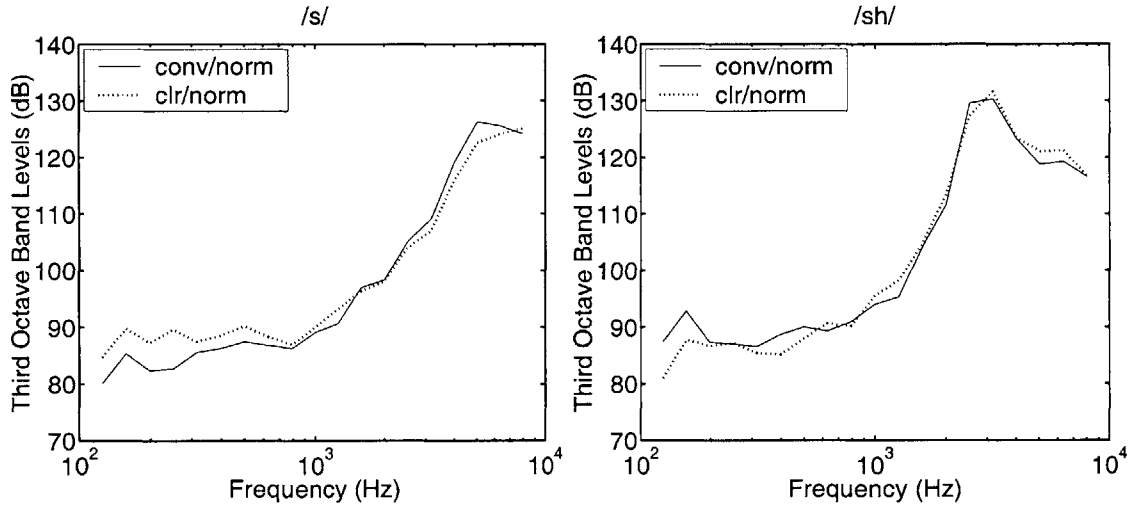

Figure 3-13: Third-octave average spectra of /s/ and /sh/ in conv/normal and clear/normal modes for SA. Similar results (no significant difference between modes) were obtained for all consonants.

Typical average vowel spectra for each talker in both modes are shown in Figures 3-14 and 3-15. The vowels displayed span the range of F1 (correlated with vowel height) and F2 (correlated with vowel fronting) and thus are fairly representative of the entire vowel space. These plots show that both talkers have higher spectral prominences at the vowel formant frequencies in clear/normal speech than in conv/normal speech. The result holds for nearly all other vowels as well (see Figures A-3 through A-10 in Appendix A). The effect appears strongest for the second and third formants (seen clearly in the spectra of individual tokens of vowels such as in Figure 3-16) and is somewhat stronger for SA than for RG. Such a spectral emphasis of formants is likely responsible for the long-term spectral differences observed in Section 3.2.3, since vowels are relatively long in duration compared to consonants and thus contribute substantially to the long-term spectrum. Moreover, the emphasis is very likely linked to the enhanced intelligibility of clear/normal speech, since vowel formants have long been considered one of the most salient cues to vowel identification[31, 43, 20]. Thus, the effect on intelligibility of increasing relative levels of vowel formants was investigated. This artificial manipulation of the signal is discussed further in Chapter 5.

3.4.4 Vowel Formant Frequencies

Vowel formant frequencies and formant bandwidths were measured using the formant tracking program provided in the *ESPS/waves+* software package. The first three formants and their bandwidths were extracted at the midpoint of each vowel and averaged over all occurrences of the vowel within a speaking style. Paired t-tests were employed to compare statistically the formants and formant bandwidths in conv/normal and clear/normal speaking styles. For formant bandwidth comparisons, the difference variable was $D = BW_{conv} - BW_{clear}$, where BW_{conv} was the formant bandwidth of the phone spoken conversationally, and BW_{clear} was the formant bandwidth of the phone spoken clearly. For formant frequency comparisons, the difference variable was $D = F_{conv} - F_{clear}$, where F_{conv} was the formant frequency of the phone spoken conversationally, and F_{clear} was the formant frequency of the phone spoken clearly.

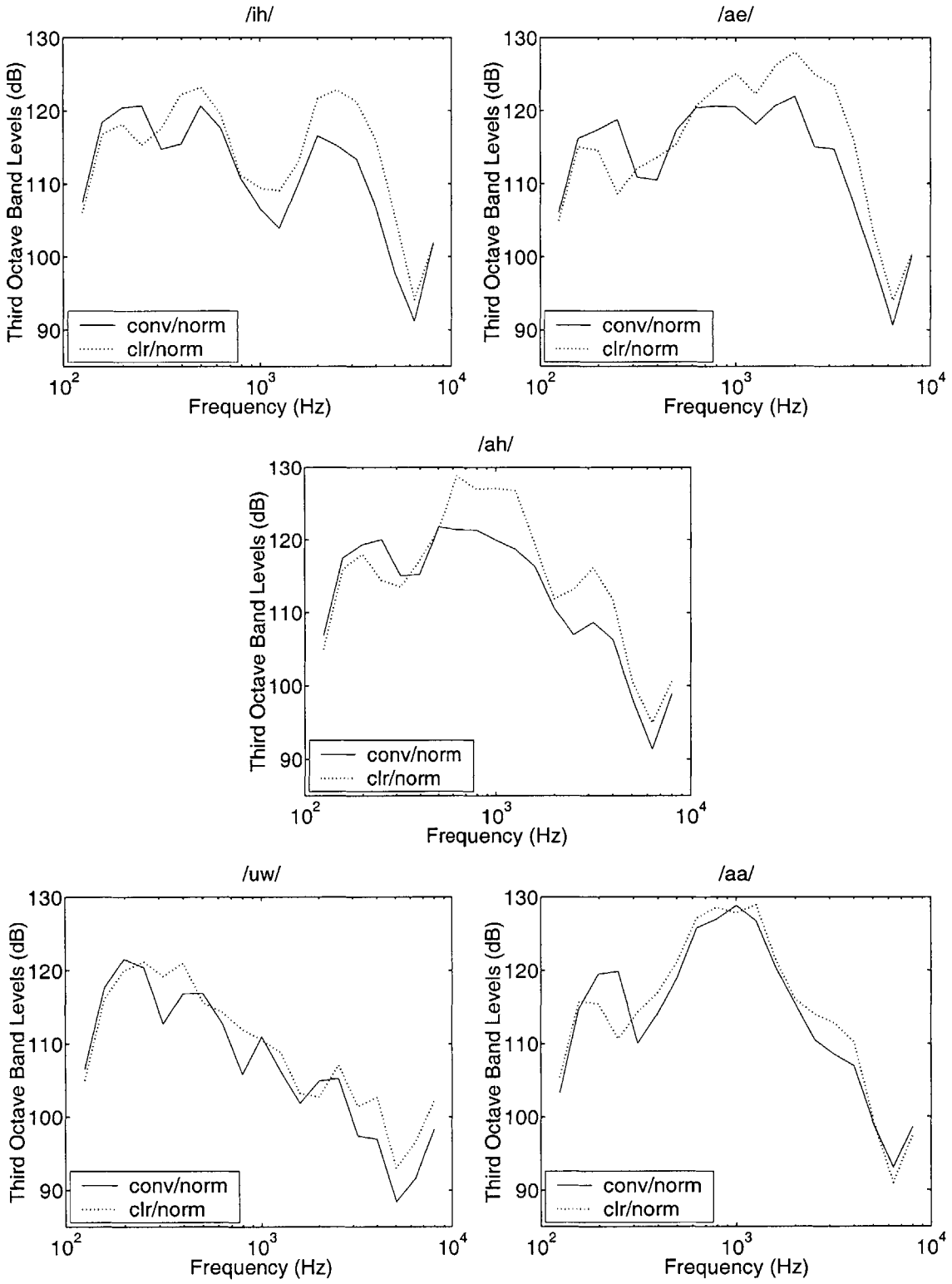

Figure 3-14: Third-octave average spectra for five vowels (spanning F1-F2 space) in conv/normal and clear/normal modes for RG. Similar results were obtained for most vowels. Spectra for remaining vowels can be found in Appendix A.

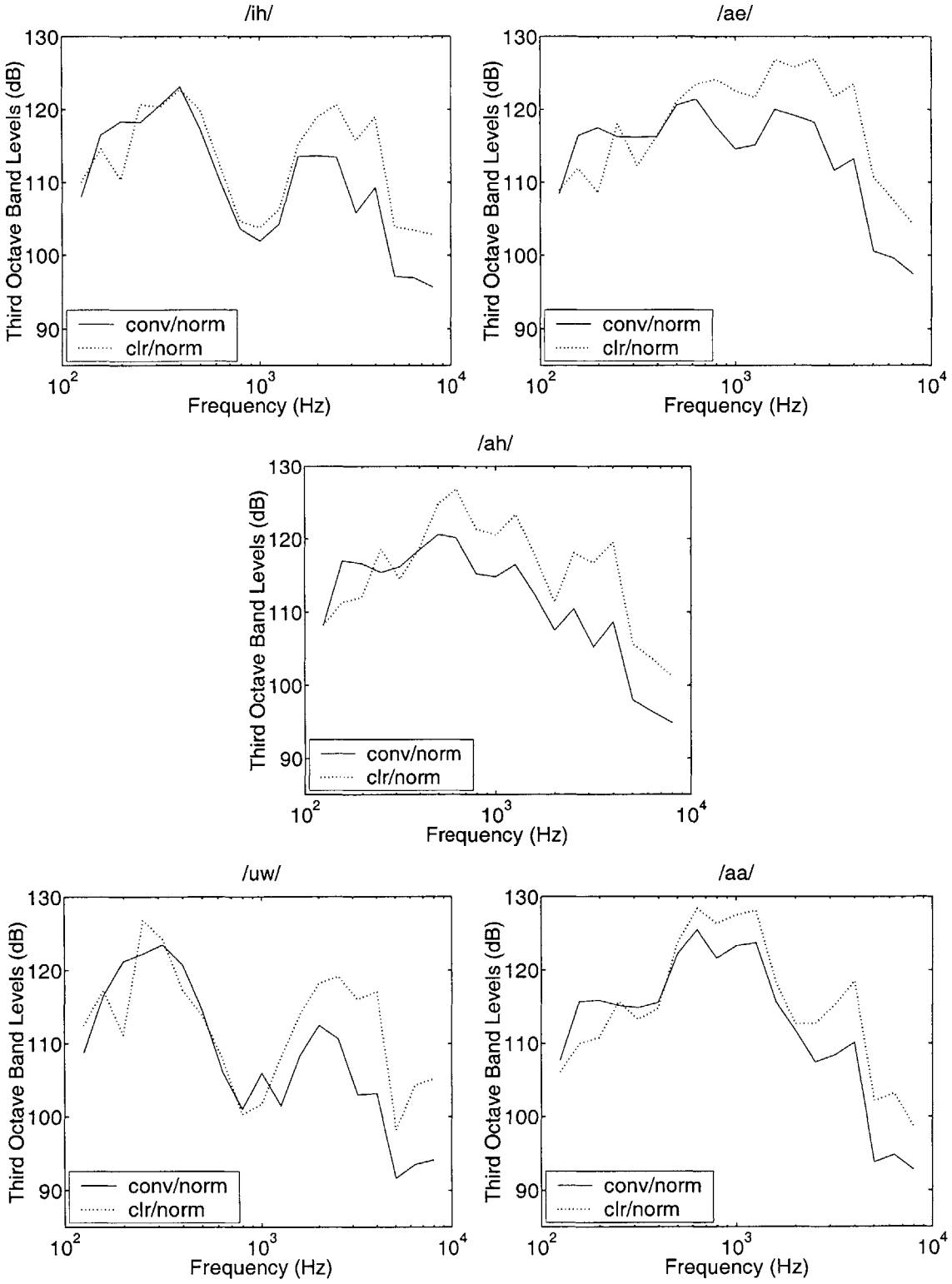

Figure 3-15: Third-octave average spectra for five vowels (spanning F1-F2 space) in conv/normal and clear/normal modes for SA. Similar results were obtained for most vowels. Spectra for remaining vowels can be found in Appendix A.

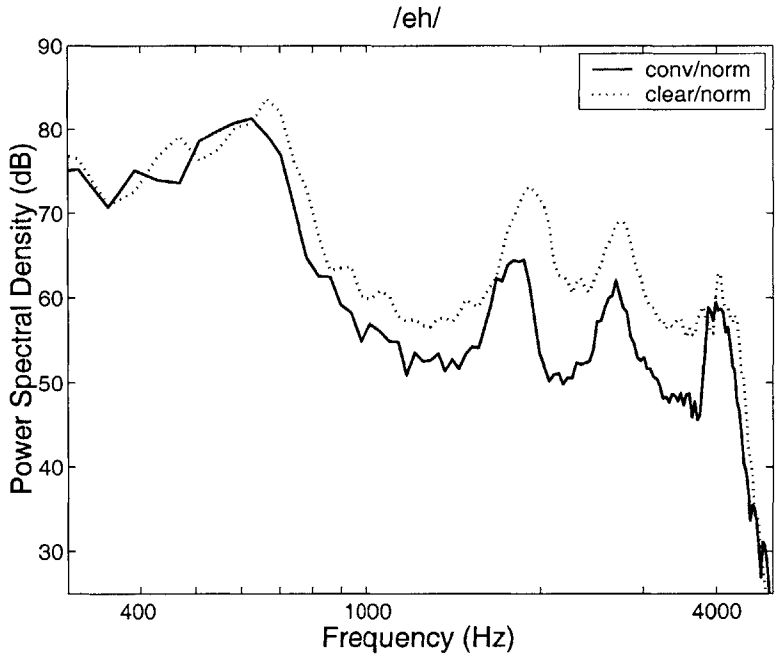

Figure 3-16: Power spectral density of an individual token (/eh/) spoken in conv/normal and clear/normal modes by SA shows relatively more power concentrated near F2 and F3 in clear/normal speech.

Formant bandwidths that differed significantly (p=0.05) between speaking modes are listed in Tables 3.9 and 3.10. Of the bandwidths that changed, all were slightly narrower in clear/normal speech than in conv/normal speech. One potential benefit of narrower bandwidths, particularly in conjunction with higher formant amplitudes (see Section 3.4.3), is that more energy is concentrated very near the vowel formant frequencies, which may aid listeners in identifying vowels.

Table 3.9: Formant bandwidth data (means in Hz) for RG in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Vowel-Formant	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	υ	Sig Level
ER-BW1	268.4	196.3	72.0	123.6	22.2	0.264	0.151	3.25	30	0.003
IH-BW1	205.2	148.5	56.7	113.6	16.9	0.452	0.002	3.35	44	0.002
OW-BW1	194.2	116.4	77.8	87.3	17.8	0.458	0.024	4.36	23	0.000
UW-BW3	195.7	134.7	61.0	56.2	22.9	0.903	0.014	2.66	5	0.045

Table 3.10: Formant bandwidth data (means in Hz) for SA in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Vowel Formant	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	υ	Sig Level
AH-BW1	296.5	213.2	83.3	155.4	22.9	0.387	0.008	3.64	45	0.001
AH-BW2	149.4	114.9	34.5	69.8	10.3	0.208	0.165	3.36	45	0.002
AO-BW1	290.5	215.2	75.3	92.3	23.8	0.670	0.006	3.16	14	0.007
AO-BW3	227.5	131.9	95.5	162.2	41.9	0.356	0.193	2.28	14	0.039
EH-BW1	203.2	176.5	26.7	71.5	11.9	0.827	0.000	2.24	35	0.031
EH-BW2	101.0	71.8	29.2	27.2	4.5	0.368	0.027	6.43	35	0.000
EH-BW3	214.2	143.6	70.6	126.8	21.1	0.362	0.030	3.34	35	0.002
EY-BW2	97.0	86.3	10.7	27.7	5.1	0.577	0.001	2.08	28	0.047
ER-BW1	185.1	142.0	43.2	73.9	16.5	0.473	0.035	2.61	19	0.017
ER-BW3	263.8	101.8	162.0	226.0	50.5	0.467	0.038	3.21	19	0.005
IH-BW2	98.9	80.4	18.5	37.3	5.0	0.756	0.000	3.72	55	0.000
IY-BW3	241.4	155.4	86.0	103.6	27.7	0.301	0.296	3.11	13	0.008

Formant frequencies that differed significantly (p=0.05) are listed in Tables 3.11 and 3.12. Although at least one formant frequency changed significantly for both talkers in over half the fifteen vowels measured, there is little in the data to suggest a trend. Very few (2–3) second or third formants exhibited a statistically significant frequency change from conv/normal to clear/normal speech for either talker, and none of these were significant for both talkers. Slightly more (3–6) first formants exhibited a statistically significant frequency change between modes for each talker. In the majority of these cases where F1 differed significantly between modes, a higher frequency was observed for clear/normal speech than for conv/normal speech. This result is consistent with the fact that clear/normal speech was produced at higher intensities than conv/normal speech, since louder speech is typically obtained with a larger jaw opening, resulting in decreased tongue height and increased F1[24].

Table 3.11: Formant frequency data (means in Hz) for RG in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Vowel-Form	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	υ	Sig Level
AA-F2	1421.9	1349.9	72.0	155.1	30.4	0.807	0.000	2.37	25	0.026
AO-F1	803.1	558.3	244.9	261.0	92.3	0.315	0.448	2.65	7	0.033
AY-F2	1419.2	1556.1	-136.9	184.0	40.1	0.418	0.059	-3.41	20	0.003
EH-F3	2939.5	2868.2	71.3	167.5	30.1	0.767	0.000	2.37	30	0.024
ER-F1	421.1	487.8	-66.7	112.1	20.1	0.340	0.062	-3.31	30	0.002
IH-F1	384.2	447.9	-63.7	107.4	16.0	0.488	0.001	-3.97	44	0.000
OW-F2	1402.7	1246.0	156.7	117.0	23.9	0.939	0.000	6.56	23	0.000
UH-F3	2688.2	2574.5	113.7	100.6	30.3	0.851	0.001	3.75	10	0.004

Since substantial changes in formant frequencies were not observed, one hypothesis is that there was a tighter clustering of formants in clear/normal speech relative to conv/normal speech, a phenomenon observed by Chen[6] in clear/slow speech. To test this hypothesis, a t-test was employed, using the test statistic outlined in[14] for

Table 3.12: Formant frequency data (means in Hz) for SA in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.05 level.

Vowel-Form	Conv Mean	Clear Mean	Mean Diff	Std Dev	Std Err	Corr	Prob	t	v	Sig Level
AE-F2	1847.7	1938.8	-91.1	108.3	25.5	0.850	0.000	-3.57	17	0.002
AH-F1	503.6	571.3	-67.7	160.4	23.7	0.457	0.001	-2.86	45	0.006
AH-F2	1372.0	1574.9	-202.9	452.9	66.8	0.272	0.067	-3.04	45	0.004
AO-F1	510.7	554.1	-43.5	70.9	18.3	0.754	0.001	-2.37	14	0.032
AW-F1	715.7	805.6	-89.9	96.9	25.0	0.834	0.000	-3.60	14	0.003
AY-F1	676.0	635.0	-60.0	67.3	15.1	0.236	0.317	-3.99	19	0.001
AY-F3	2400.7	2620.6	-220.0	275.1	61.5	0.664	0.001	-3.58	19	0.002
EH-F2	1722.8	1784.4	-61.6	101.3	16.9	0.886	0.000	-3.65	35	0.001
EY-F1	487.2	518.8	-31.6	74.0	13.7	0.631	0.000	-2.30	28	0.029
EY-F3	2617.0	2706.1	-89.1	196.6	36.5	0.336	0.074	-2.44	28	0.021
IY-F1	309.5	282.6	26.9	41.9	11.2	0.285	0.323	2.40	13	0.032

variances associated with paired observations:

$$t = \frac{s_1^2 - s_2^2}{2s_1 s_2 \sqrt{\frac{1-r_{12}^2}{v}}}$$

where s_1^2 and s_2^2 are the sample variances of conv/normal and clear/normal formants, respectively, r_{12} is the correlation coefficient between paired observations, and v represents the degrees of freedom.

Tables 3.13 and 3.14 show formant frequency variances that differed significantly (alpha=0.05) between speaking styles. Slightly less than one-fourth of the formant frequencies exhibited a statistically significant change in variance. Of these, just over half showed a variance increase from conv/normal to clear/normal modes. Thus, there is not a trend in the data indicating that formant frequencies may cluster more tightly in clear/normal speech. A stronger increase in clustering of formant frequencies for tense vowels was previously reported for clear/slow speech by Chen[6]. The differences between Chen's data and this study are most likely a result of Chen's use of CVC materials, which introduces less contextual variability into the vowels than sentence

materials. Formant frequency clustering was also examined for vowels in key words and in word-initial positions. Results were very similar and can be found in Tables A.3 through A.6 in Appendix A.

Table 3.13: Formant frequency variance data (means in Hz) for RG in conversational and clear modes at normal rate. Table shows only cases where a change in variance for clear/normal speech was significant in paired t-tests at the alpha=0.05 level. N indicates number of cases.

Phone	Formant	StDev Diff	Conv Mean	Conv StDev	Clear Mean	Clear StDev	N	Corr	Alpha
AH	F3	141	2771	440	2872	299	27	0.231	0.027
AA	F3	-76	2569	307	2498	383	26	0.849	0.024
AH	F2	-161	1468	237	1612	398	27	0.348	0.004
AX	F3	-71	2953	130	3000	201	11	0.737	0.039
AY	F3	-67	2792	141	2744	208	21	0.656	0.017
EH	F2	-54	1988	173	1943	227	31	0.801	0.010
ER	F1	39	421	113	488	74	31	0.340	0.009
IH	F1	24	384	116	448	92	45	0.488	0.045
UH	F3	-78	2688	84	2574	162	11	0.851	0.002
UW	F1	-91	346	70	428	161	6	0.904	0.006

In order to facilitate further comparison with previous work, plots comparing the first and second formants of conv/normal and clear/slow speech are shown in Figure 3-17. Since results for the third formant were similar, discussion here is limited to the first two formants. As in previous studies[46, 6], the vowel space for clear/slow speech is expanded relative to conv/normal speech. The expanded vowel space, however, is less evident in clear/normal speech (see Figure 3-18). It can be seen only for tense vowels produced by SA. Again, there is little indication (the strongest trend in RG's data for clear/slow speech is still fairly weak) that formant frequencies in clear speech cluster more tightly than in conversational speech, which was reported previously by Chen[6] for CVC contexts. The data in this study, however, are similar to results for clear/slow speech were reported by Picheny et al.[46], who also used sentence materials.

In summary, vowel formants extracted at the midpoints of vowels do not appear to cluster more tightly or to be significantly closer to targets in clear/normal speech than

Table 3.14: Formant frequency variance data (means in Hz) for SA in conversational and clear modes at normal rate. Table shows only cases where a change in variance for clear/normal speech was significant in paired t-tests at the alpha=0.05 level. N indicates number of cases.

Phone	Formant	StDev Diff	Conv Mean	Conv StDev	Clear Mean	Clear StDev	N	Corr	Alpha
AA	F2	-222	1384	300	1465	522	17	0.453	0.012
AH	F1	-39	504	131	571	170	46	0.457	0.027
AH	F2	-216	1372	239	1575	455	46	0.272	0.000
AO	F2	-108	1094	156	1162	264	15	0.373	0.026
EH	F1	27	519	98	535	71	36	0.666	0.006
EH	F2	-72	1723	117	1784	189	36	0.886	0.000
EY	F1	33	487	95	519	62	29	0.631	0.003
EY	F2	-56	1871	135	1920	191	29	0.635	0.012
IH	F3	80	2642	224	2693	144	56	0.253	0.001
UW	F2	-86	1834	212	1818	298	9	0.898	0.038

in conv/normal speech. While it was hypothesized that formant frequencies would change in a way that would emphasize or exaggerate articulatory targets in clear/normal speech, it cannot generally be seen from the data. One possible explanation for this deviation from the theoretical prediction is that the formants were measured only once, at the vowel midpoint. There is growing evidence that listeners rely on formant frequency movements throughout the vowel and not just at CV and VC transitions (e.g.,[21]). It is possible that formant contours in clear/normal speech more closely matched target values, but cannot be sufficiently described by a measurement of the formant frequencies at one point in time. In any case, changes in steady-state vowel formant locations are not likely to account for the high intelligibility of clear/normal speech.

3.4.5 Formant Transition Duration and Extent

When duration and extent of the formant transitions were increased, Ochs *et al.*[40] found that both normal hearing and hearing-impaired listeners showed improved ability to identify stop consonants. In addition, Tallal *et al*[51, 52] showed that increas-

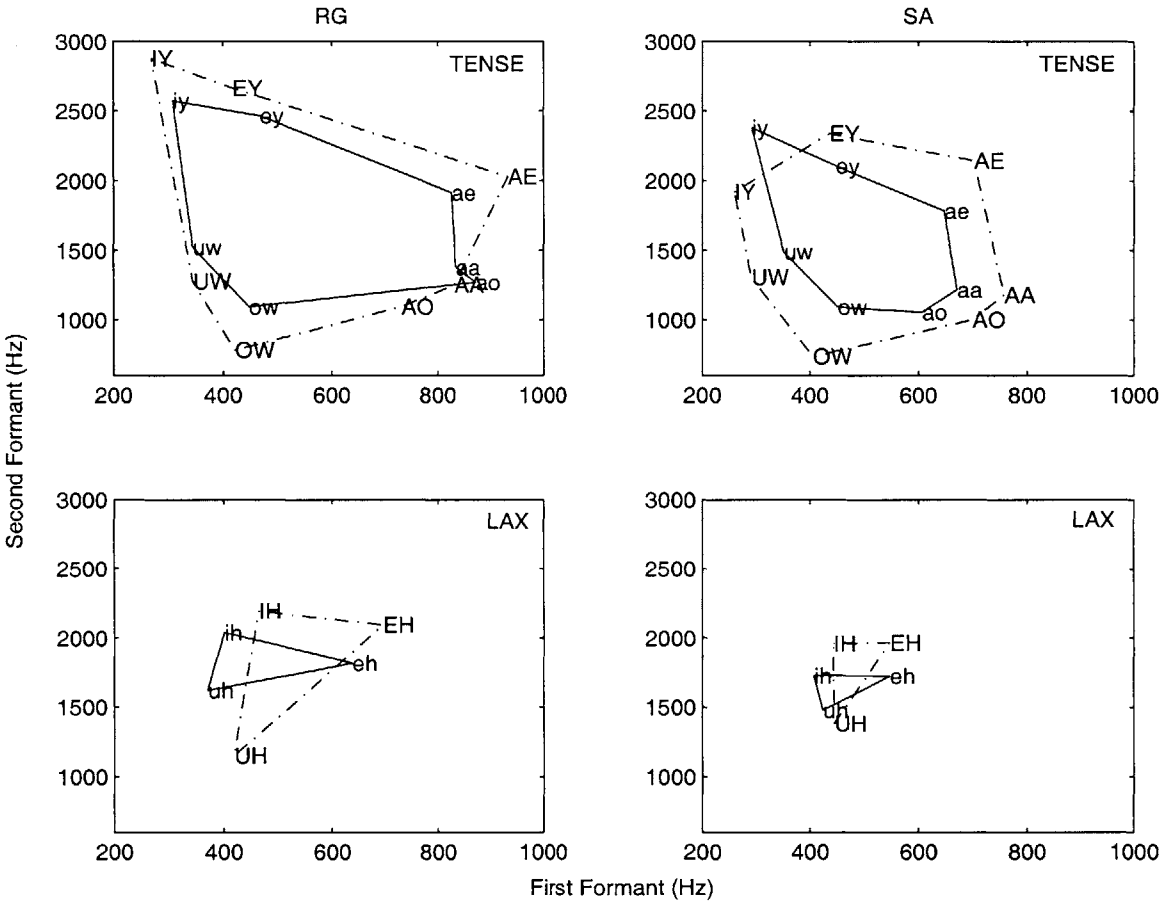

Figure 3-17: Vowel formant frequency data. Top row shows results for tense vowels and bottom row for lax vowels; columns give results for each talker. Conv/slow speech is indicated by lower case phone labels and solid lines, and clear/slow speech is indicated by upper case phone labels and dash-dotted lines.

ing formant transition duration, but not extent, led to improvement in the ability of language-learning impaired children to identify speech sounds. However, Turner *et al.*[53] found that when extent is held constant, increasing transition duration helped only normal hearing listeners and provided little benefit to hearing-impaired listeners. These studies indicate that formant transition duration and extent may play an important role in speech intelligibility. Therefore, both factors were examined in clear/normal and conv/normal speech.

Formant transition data were collected from the database for CV sequences beginning with stop consonants (/p/, /t/, /k/, /b/, /d/, and /g/), since the start

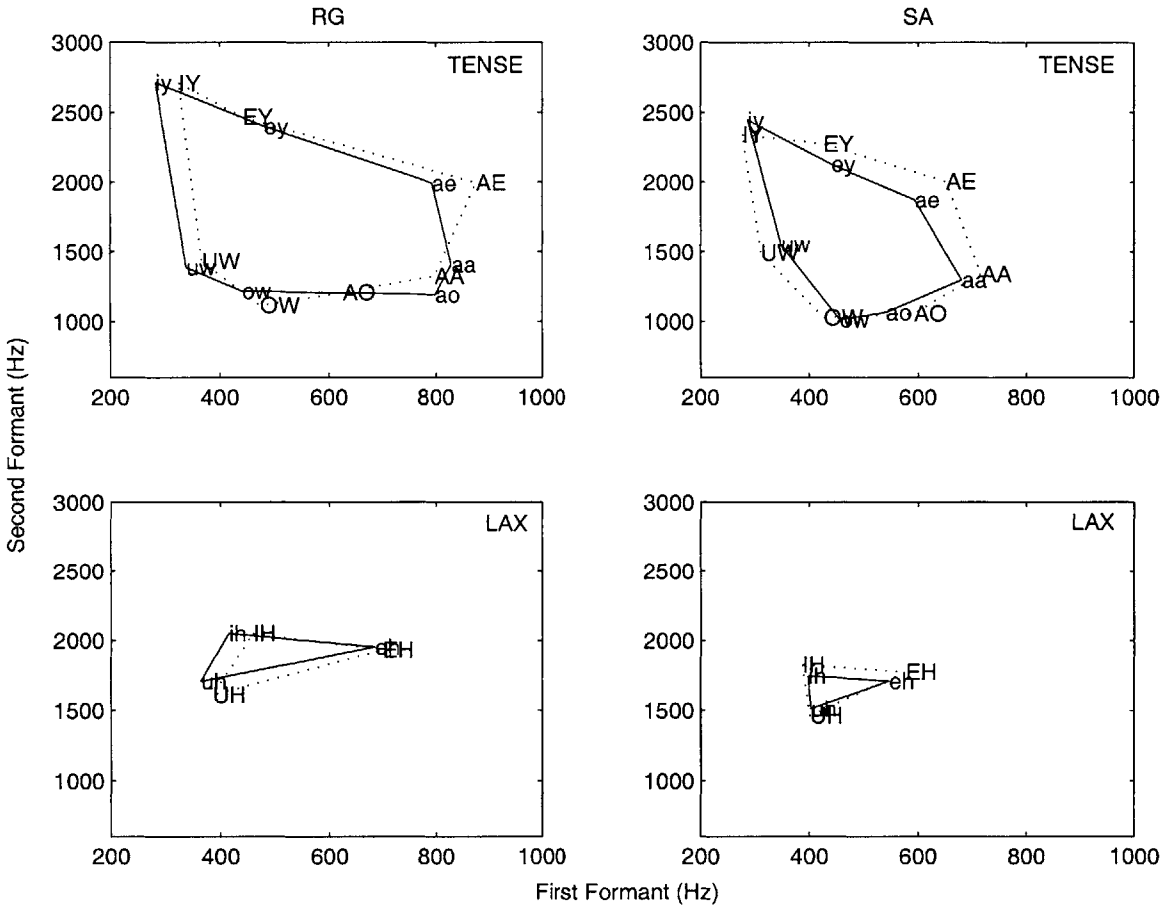

Figure 3-18: Vowel formant frequency data. Top row shows results for tense vowels and bottom row for lax vowels; columns give results for each talker. Conv/normal speech is indicated by lower case phone labels and solid lines, and clear/normal speech is indicated by upper case phone labels and dotted lines.

of the formant transition for these consonants is relatively well-defined. A semi-automated procedure was developed to tabulate the formant transition data. Using the *ESPS/waves+* software package, a human labeller marked the beginning and end of each transition using the time waveform and the spectrogram to determine when the vowel had reached steady-state and the formant transition was complete. The value of the first three formants at the beginning and end of the transition as well as the length of the transition were then automatically recorded. Finally, the duration, rate, beginning and ending frequency values, and extent of the formant transitions found in conv/normal and clear/normal styles were compared statistically

via paired t-tests. With only about 100 CV tokens beginning with stop consonants in the database, there were not many instances of each CV pair to be analyzed. Therefore, the consonants and vowels were each grouped according to place of articulation. Stop consonants were categorized as labial (/p/ and /b), alveolar (/t/ and /d/), and velar (/k/ and /g/). Vowels were classified into HighFront (/ih/ and /iy/), MidFront (/eh/ and /ey/), LowFront (/ae/), Neutral (/ah/ and /ax/), HighBack (/uh/ and /uw/), and LowBack (/aa/, /ao/, and /ow/). Variables that differed significantly (p=0.1) are listed in Tables 3.15 and 3.16. In no case did the duration of the formant transition change significantly between conv/normal and clear/normal speech. Moreover, transition rate and transition extent changed significantly in only about three cases for each of the talkers. However, even when consonants and vowels were grouped by place of articulation, there were very few data points for each CV pair analyzed. Therefore, these results are somewhat inconclusive. This is evidenced by the preponderance of end formant frequency values that have been recognized as significant. Since these values are nominally the steady-state values, they should be reasonably close to the formant values measured at the vowel midpoint. In Section 3.4.4, vowel midpoint frequencies were not found to change significantly.

Table 3.15: Formant transition data for RG in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.1 level.

Cons-Vowel	Var	Cnv Mean	Clr Mean	Mean Diff	St Dv	St Er	Corr	Prob	t	υ	Sig Level
Labial-MidFront	F3-BEG	3005 (Hz)	2816 (Hz)	188.6	104	39	0.949	0.001	4.79	6	0.003
Labial-MidFront	F3-EXT	62 (Hz)	184 (Hz)	-122.1	161	61	0.905	0.005	-2.00	6	0.092
Labial-LowFront	F1-EXT	229 (Hz)	322 (Hz)	-93.7	48	28	0.999	0.028	-3.38	2	0.077
Alveolar-HighFront	F2-EXT	183 (Hz)	301 (Hz)	-117.7	156	59	0.761	0.047	-2.00	6	0.092
Alveolar-MidFront	F1-END	715 (Hz)	611 (Hz)	104.4	55	25	0.957	0.011	4.24	4	0.013
Alveolar-LowFront	F1-END	534 (Hz)	684 (Hz)	-150.5	22	16	1.000	0.000	-9.71	1	0.065
Alveolar-LowBack	F1-END	761 (Hz)	694 (Hz)	66.3	83	28	0.913	0.001	2.40	8	0.043
Velar-LowBack	F2-END	1331 (Hz)	1143 (Hz)	189.1	208	79	0.656	0.110	2.41	6	0.053
Velar-LowBack	F3-END	2745 (Hz)	2546 (Hz)	199.0	134	51	0.879	0.009	3.92	6	0.008

Table 3.16: Formant transition data for SA in conversational and clear modes at normal rate. Table shows only phones that were significant in paired t-tests at the p=0.1 level.

Cons-Vowel	Var	Cnv Mean	Clr Mean	Mn Diff	St Dv	St Er	Corr	Prob	t	v	Sig Level
Labial-HighFrnt	F1-END	373 (Hz)	330 (Hz)	43	4	2	0.999	0.028	17.75	2	0.003
Labial-HighFrnt	F1-SPD	5082 (Hz/s)	2090 (Hz/s)	2991	426	246	0.985	0.111	12.16	2	0.007
Labial-MidFrnt	F3-END	2535 (Hz)	2647 (Hz)	-112	48	24	0.934	0.066	-4.71	3	0.018
Labial-LowFrnt	F2-END	1857 (Hz)	2002 (Hz)	-146	129	58	0.957	0.011	-2.52	4	0.065
Alveolar-HighFrnt	F2-BEG	1809 (Hz)	1893 (Hz)	-85	101	34	0.902	0.001	-2.51	8	0.037
Alveolar-HighFrnt	F2-END	1897 (Hz)	2040 (Hz)	-143	78	26	0.960	0.000	-5.50	8	0.001
Alveolar-HighFrnt	F3-END	2688 (Hz)	2752 (Hz)	-64	48	16	0.872	0.002	-4.00	8	0.004
Alveolar-HighFrnt	F2-EXT	73 (Hz)	146 (Hz)	-74	115	35	0.681	0.021	-2.12	10	0.060
Alveolar-MidFrnt	F2-END	1786 (Hz)	1873 (Hz)	-87	77	29	0.890	0.007	-3.01	6	0.024
Alveolar-MidFrnt	F3-END	2680 (Hz)	2725 (Hz)	-45	45	17	0.669	0.101	-2.66	6	0.037
Alveolar-LowBack	F1-END	555 (Hz)	903 (Hz)	-38	12	5	0.997	0.000	-7.46	5	0.001
Velar-HighFrnt	F3-END	2728 (Hz)	2829 (Hz)	-101	81	40	0.378	0.622	-2.51	3	0.087
Velar-MidFrnt	F3-EXT	-25 (Hz)	-253 (Hz)	228	4	3	1.000	0.000	76.00	1	0.008
Velar-MidFrnt	F3-SPD	-710 (Hz/s)	-5070 (Hz/s)	4360	281	199	1.000	0.000	21.91	1	0.029
Velar-LowFrnt	F1-END	279 (Hz)	405 (Hz)	-126	68	39	0.987	0.104	-3.20	2	0.085
Velar-Neutral	F3-END	2472 (Hz)	2619 (Hz)	-147	93	42	0.800	0.104	-3.53	4	0.024

3.4.6 Consonant-Vowel Ratio

Much attention in the literature (e.g.,[37, 15] has been given to consonant-vowel ratio (CVR), or ratio of consonant power to the power in the nearest adjacent vowel. Picheny et al.[46] did not measure consonant-vowel ratio per se, instead criticizing the utility of CVR measurements in sentential environments. That study did, however, report an increase in the relative power of some, particularly unvoiced, consonants. Since such an increase in power was not observed for clear/normal speech (see Section 3.4.1), CVR was also measured for the sake of completeness.

Consonants were grouped into five classes: stops, fricatives, affricates, liquids, and, nasals. For stop consonants, the power was designated as the power in the stop burst. Affricate power was defined as the non-closure power. For all other consonant classes and for vowels, the rms level was used. Resulting CVRs are plotted in Figures 3-19 and 3-20. Little trend is evident between speaking styles for any of the consonant classes. Highly intelligible speech does not appear to result from consistent changes in CVR.

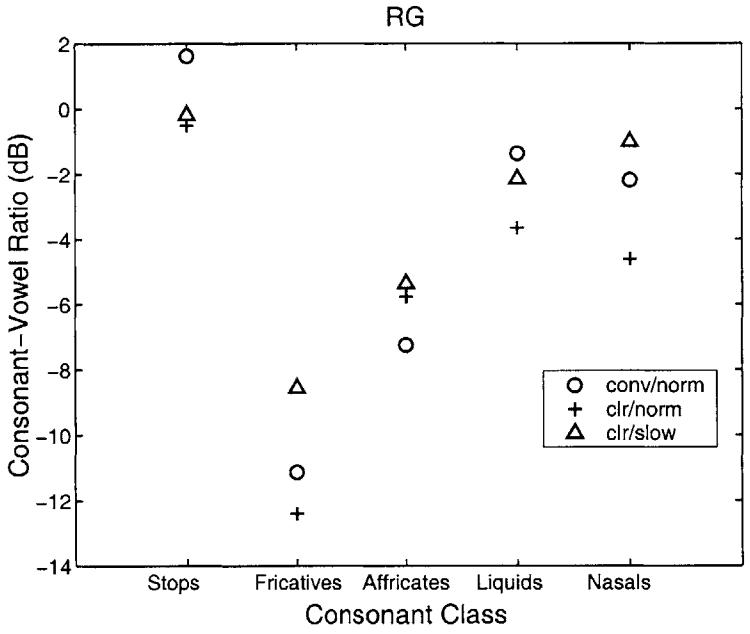

Figure 3-19: Comparison of conv/normal, clear/normal, and clear/slow consonant-vowel ratios for RG.

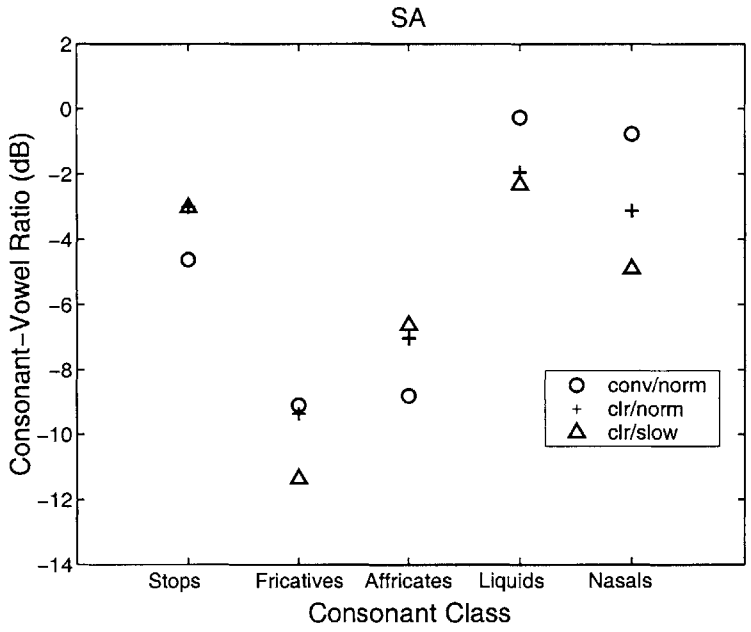

Figure 3-20: Comparison of conv/normal, clear/normal, and clear/slow consonant-vowel ratios for SA.

3.4.7 Voice-onset Time

Voice-onset time (VOT) is widely considered an important cue for voicing contrasts in stop consonant identification (e.g., [31, 13]). Therefore, VOTs were measured for word-initial stop consonants in the database using a semi-automated procedure similar to that used for formant data tabulation. Using the *ESPS/waves+* software package, a human labeller marked the release of the stop burst and the onset of voicing with the mouse, relying on information available in the time waveform and the spectrogram to judge when voicing had begun. The VOT was then automatically calculated and recorded. Results comparing VOTs in clear/normal speech to VOTs in conv/normal speech are shown in Figure 3-21. The overall results are in good agreement with average VOTs previously reported for stop consonants, showing clearly that unvoiced stops have longer VOTs and that VOT increases as place of articulation moves from labial to to alveolar to velar.

Changes in VOT between conv/normal and clear/normal speech appear primarily in voiceless stop consonants. Unfortunately, the talkers once again exhibit opposite

trends in the data. The trend in SA's data reveals an increase in VOT in clear/normal speech, while RG's data shows a decrease. This difference is most likely another indication of different talker strategies for speaking clearly at normal rates and is discussed further in Chapter 4.

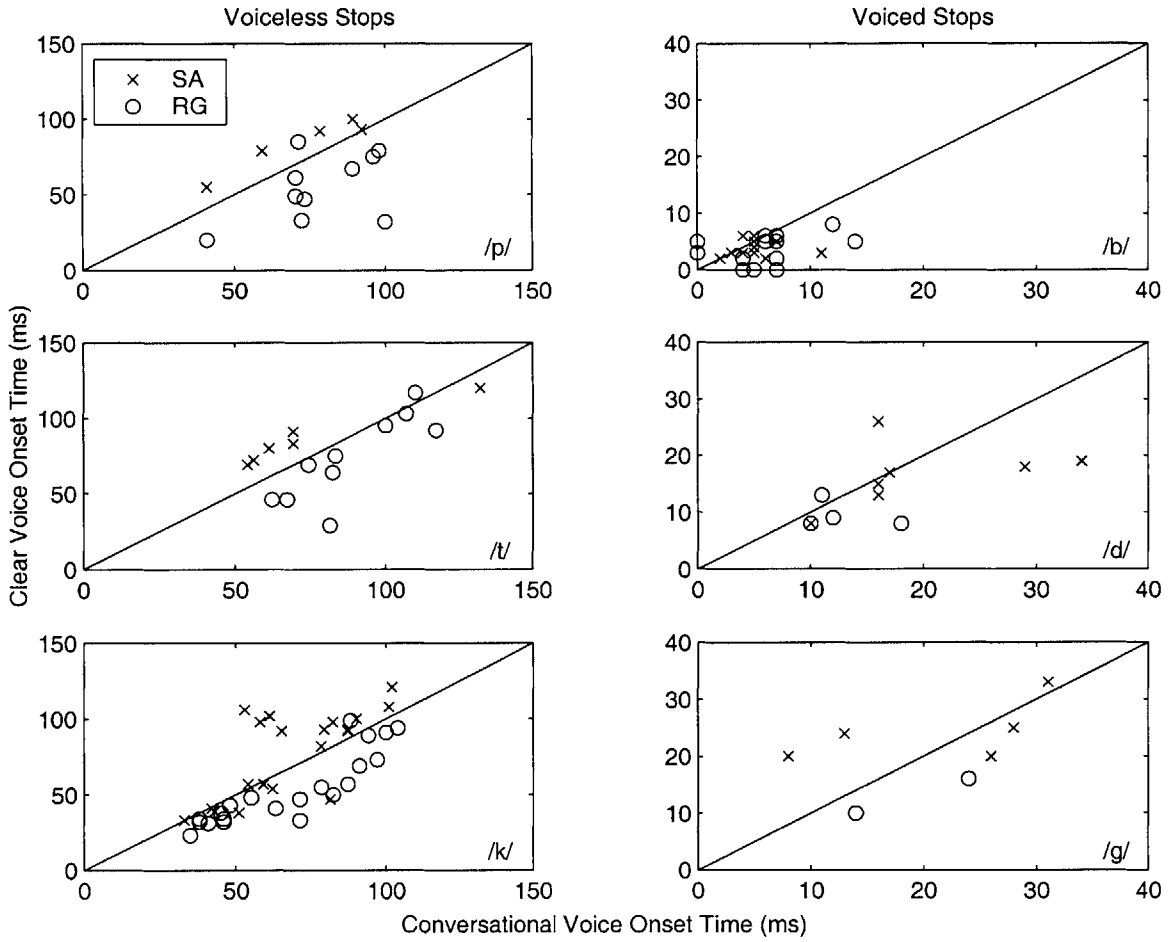

Figure 3-21: Comparison of stop consonant voice-onset time in clear and conversational modes at normal speaking rates.

3.5 Summary

In this study, differences in acoustic characteristics of clear/slow speech relative to conv/normal speech are consistent with previously reported results. Many of these differences, however, are not apparent when comparing clear/normal speech to conv/nor-

mal speech. Since talkers presumably do not have time to retain all of the characteristics of clear/slow speech when speaking at normal rates, they must pick a subset of characteristics to continue to emphasize in clear/normal speech.

One difference between clear/normal and conv/normal speech was evident for both talkers, an increase in energy near vowel formant frequencies. The short-term vowel spectra showed that both talkers raised the relative amplitude of formants in clear/normal speech. Such a spectral emphasis on formants is likely responsible for the long-term spectral differences observed as well, since vowels are relatively long in duration compared to consonants and thus contribute substantially to the long-term spectrum. This relative energy increase near formants is very likely linked to the enhanced intelligibility of clear/normal speech, since vowel formants have long been considered one of the most salient cues to vowel identification.

For several other acoustic measurements, however, the changes between clear/normal and conv/normal speech differed dramatically between talkers. For example, RG released more stop bursts in clear/normal speech, while SA actually released fewer. SA, however, preserved several other characteristics of clear/slow speech in his clear/normal speech. First, his F0 average and range increased. Second, SA showed a greater modulation depth for slow modulation frequencies (less than 3–4Hz) in most octave bands. Third, VOTs for SA's stop consonants were longer in clear/normal speech than in conversational speech. These differences are most likely a result of different talker strategies for producing clear speech at normal rates. This phenomenon is analyzed in more depth in Chapter 4.

Chapter 4

Assessment of Acoustical Results

Fewer acoustical differences were observed between conv/normal and clear/normal speech than between conv/normal and clear/slow speech. Nonetheless, since clear/normal speech exhibits a significant intelligibility advantage over conversational speech, it can be assumed that at least some of the essential acoustical characteristics of clear speech are preserved in clear/normal speech. However, since the the differences reported in Chapter 3 are not dramatic, there was some concern that these differences may not be able to account for a large part of the intelligibility advantage found in clear/normal speech. Therefore, the acoustical results were re-evaluated in two different contexts.

First, the acoustic data were analyzed with multivariate statistical techniques, since one explanation of the observed acoustic characteristics is that the essential characteristic of clear speech may be a linear combination of several acoustical parameters. In this case, it is possible that comparing individual acoustic characteristics could prove inadequate for capturing the difference between speaking modes, which would account for the fact that fewer individual acoustic differences were found between conv/normal and clear/normal speech.

Second, additional intelligibility tests were performed to investigate whether each talker may have employed a somewhat different strategy for producing clear/normal speech, since some of the measured acoustic differences between clear/normal and conv/normal speech (e.g. segment duration and VOT) differed dramatically between

talkers. Thus, when producing clear/normal speech, each talker may have preserved different subsets of the acoustical characteristics found in clear/slow speech. In this case, each talker's clear/normal speech could be considered a different type of clear speech. Furthermore, even though these types of clear speech provide an intelligibility advantage for normal hearing listeners in noise, it seems likely that this advantage could vary in other degraded environments.

4.1 Multivariate Statistical Analysis

Several multivariate statistical techniques were reviewed, and it was concluded that discriminant analysis was best suited to the data in this study. Discriminant analysis is a technique for attempting to distinguish two groups of data (in this case, conv/normal and clear/normal speech) based on discriminating variables, or measured characteristics of the data that are expected to differ. The analysis derives a linear combination of the variables called the "discriminating function" to maximize the separation of the groups. The discriminating function is of the form

$$D = d_1 Z_1 + d_2 Z_2 + ... + d_p Z_p$$

where D is the score on the discriminant function, the ds are the weighting coefficients, and the Zs are the standardized values of the p discriminating variables.

For both talkers, discriminant analysis was performed on data for classes of phonemes, with the vowel class being broken into subclasses based on place of articulation: HighFront (/ih/ and /iy/), MidFront (/eh/ and /ey/), LowFront (/ae/), Neutral (/ah/ and /ax/), HighBack (/uh/ and /uw/), and LowBack (/aa/, /ao/, and /ow/). The results are shown in Tables 4.1 and 4.2. The technique was largely unsuccessful in discriminating RG's conv/normal and clear/normal tokens, as can be seen by the fact that the smallest Wilk's Lambda (a measure of the variance not explained by the discriminating function) obtained for any phoneme group was 0.45. On the other hand, the analysis did discriminate fairly well between SA's conv/normal and

clear/normal vowels, with the exception of neutral vowels. Upon closer inspection, however, it seems that the most heavily weighted discriminating variable was consistently F0. Moreover, the weightings of the other variables across vowel subclass is quite variable. Thus, the discriminant analysis provided little new insight into the acoustical analysis. There do not appear to be linear functions of acoustical parameters that are variables are responsible for the differences between conv/normal and clear/normal speech.

Table 4.1: Discriminant analysis results (linear weightings of discriminating variables) for RG data. D is the duration variable and P is the power vavriable.

Phone	D	P	F0	F1	F2	F3	F4	BW 1	BW 2	BW 3	BW 4	Corr	λ
ih, iy	0.6	0.5	0	-0.7	0	-0.2	0	0.7	0.5	-0.4	0.6	0.587	0.66
eh, ey	0.3	0	0.3	-0.5	-0.6	0	0	0.3	0	0	0.8	0.416	0.83
ae	-0.5	0	0	0.9	0	0	0	0	0	0.4	-0.6	0.421	0.82
aa, ao, ow	0	0	0	-0.5	0.3	0	0	0	0.6	0.4	0.4	0.370	0.86
uh, uw	0	0.8	-0.5	-1.2	0	0.6	0	0.5	0	0	0	0.535	0.71
ah, ax	0.6	0.9	-0.3	-1.0	0	0	-0.4	0.4	0	0	0	0.373	0.86
s,sh, z,zh	0	1.0	0	0	0	0	0	0	0	0	0	0.076	0.99
w,l, r,y	0.3	0.8	0	-0.7	0.4	0	-0.3	0.9	0	0.3	0	0.439	0.81
m,n, ng	0	0.9	0	-0.2	0	0	0	0	0	0.1	0.3	0.740	0.45
ch, jh	0	0	0	0	0	0	0	0	0	0	0	n/a	n/a
cl, vcl	0.8	-0.7	0	0	0	0	0	0	0	0	0	0.191	0.96
bst	0.8	0.8	0	0	0	0	0	0	0	0	0	0.331	0.89

Table 4.2: Discriminant analysis results (linear weightings of discriminating variables) for SA data. D is the duration variable and P is the power vavriable.

Phone	D	P	F0	F1	F2	F3	F4	BW 1	BW 2	BW 3	BW 4	Corr	λ
ih, iy	0	0.9	-1.2	-0.2	0	-0.3	0.3	0	0.2	0.3	0.2	0.852	0.27
eh, ey	0.5	-0.3	1.2	0	-0.2	0	0	-0.5	-0.2	0	0	0.856	0.27
ae	0.3	-0.3	1.1	0	0	0	0	0	-0.3	0	0	0.890	0.21
aa, ao, ow	0	-0.2	1.1	0	0.3	-0.4	0.2	0	0	0	-0.3	0.828	0.31
uh, uw	0	-1.0	1.4	0.2	0	0.5	-0.3	0	0	0	-0.2	0.869	0.24
ah, ax	0.1	-0.3	1.0	0	0	0.1	0	0	0	-0.1	0	0.748	0.44
s,sh, z,zh	0	1.0	0	0	0	0	0	0	0	0	0	0.123	0.99
w,l, r,y	0.1	-0.4	1.0	0	0	-0.1	0.2	0	0	0	-0.2	0.739	0.45
m,n, ng	0.3	-0.7	1.1	0	-0.2	0	0.5	0	-0.4	-0.3	-0.2	0.849	0.28
ch, jh	0	0	0	0	0	0	0	0	0	0	0	n/a	n/a
cl, vcl	-0.4	1.0	0	0	0	0	0	0	0	0	0	0.149	0.98
bst	0	1.0	0	0	0	0	0	0	0	0	0	0.212	0.95

4.2 Evaluation of Clear Speech in Other Degraded Environments

In order to investigate further whether the two talkers may have employed different strategies when producing clear speech at normal rates as the acoustical data in Chapter 3 suggests, tests were conducted to evaluate the intelligibility of these talkers in other degraded environments. These tests also helped to assess the robustness of the high intelligibility of clear/normal speech for each talker.

4.2.1 Intelligibility Tests

Three types of degradations were presented to normal hearing listeners: reverberation, low-pass filtering, and high-pass filtering conditions. In addition, non-native listeners were tested in additive noise.

Reverberant Environment

The reverberant test environment was simulated by convolving each sentence with the impulse response of the "conference room" environment described by Payton *et al.*[42], a room with reverberation time of 0.60 seconds (the time required for the room's impulse response to decay 60dB from its maximum rms level). As in Payton *et al.*, speech-shaped noise was added to the sentences at a signal-to-noise ratio of 0dB prior to convolution.

Five normal-hearing listeners were employed to evaluate the intelligibility of the processed sentences. One hundred nonsense sentences were presented for each talker in conv/normal and conv/slow modes, and fifty sentences were produced by each talker in clear/normal and clear/slow modes. The percent-correct key word intelligibility scores for each talker, averaged across listener, are presented in Figure 4-1. At both slow and normal rates, SA achieved an intelligibility benefit in this reverberant environment by speaking clearly, whereas RG only achieved such an improvement at a slow speaking rate.

Low-Pass Environment

Using the General Radio model 1925 1/3-octave multifilter bank, a low-pass environment was approximated by passing each sentence through the 1/3-octave filter bands with center frequencies ranging from 80Hz to 1000Hz, as in condition 2 of Grant and Braida[16]. To ensure that any remaining out-of-band speech energy did not affect intelligibility, speech-shaped noise was added to the filtered speech at roughly 35dB below the peak speech level.

Another five normal-hearing listeners were employed to evaluate the intelligibi-

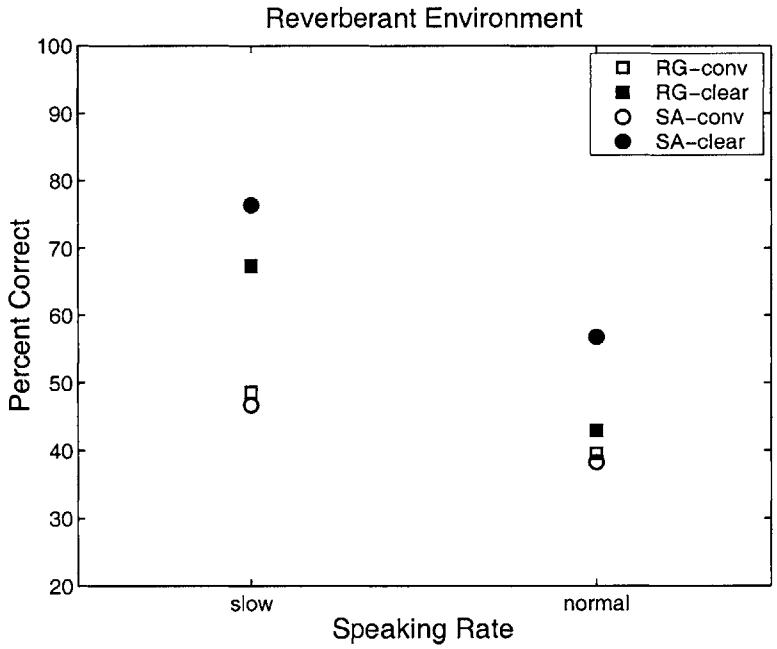

Figure 4-1: Intelligibility data versus rate in a reverberant (RT=0.6s) environment. Squares indicate RG and circles indicate SA. Open symbols are data for conversational speech, and filled symbols are data for clear speech.

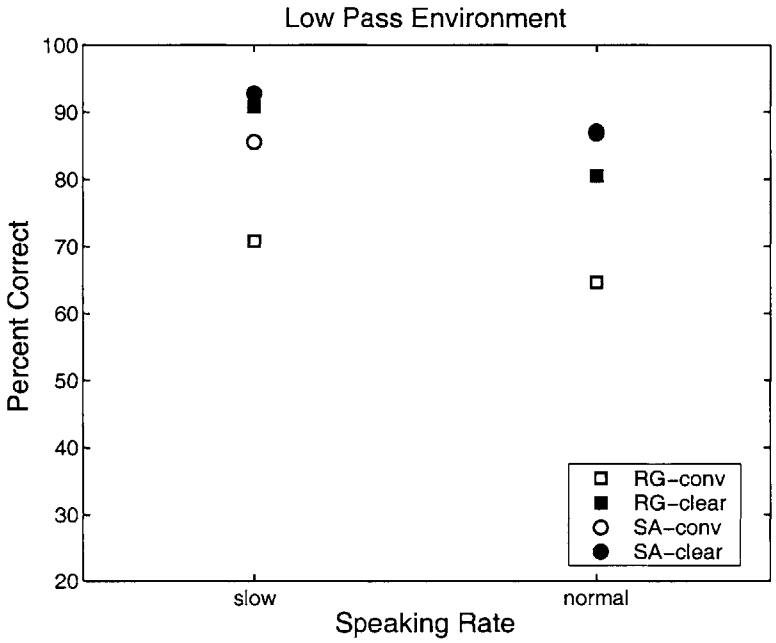

Figure 4-2: Intelligibility data versus rate in a low pass environment. Squares indicate RG and circles indicate SA. Open symbols are data for conversational speech, and filled symbols are data for clear speech.

lity of this processing. Again, one hundred nonsense sentences from each talker in conv/normal and conv/slow modes and fifty sentences from each talker in clear/normal and clear/slow modes were presented. The percent-correct key word intelligibility scores for each talker, averaged across listener, are presented in Figure 4-2. In this environment, both talkers achieved an intelligibility benefit by speaking clearly at slow rates, but only RG achieved a similar benefit at her normal speaking rate. The intelligibility of SA's conv/normal speech (87%), however, was high enough that a ceiling effect may have prevented listeners from achieving higher scores for his clear/normal speech.

High-Pass Environment

Again using the General Radio model 1925 filter bank, a high-pass environment was approximated by passing each sentence through the 1/3-octave filter centered at 3150Hz, as in condition 5 of Grant and Braida[16]. As in the low pass-environment,

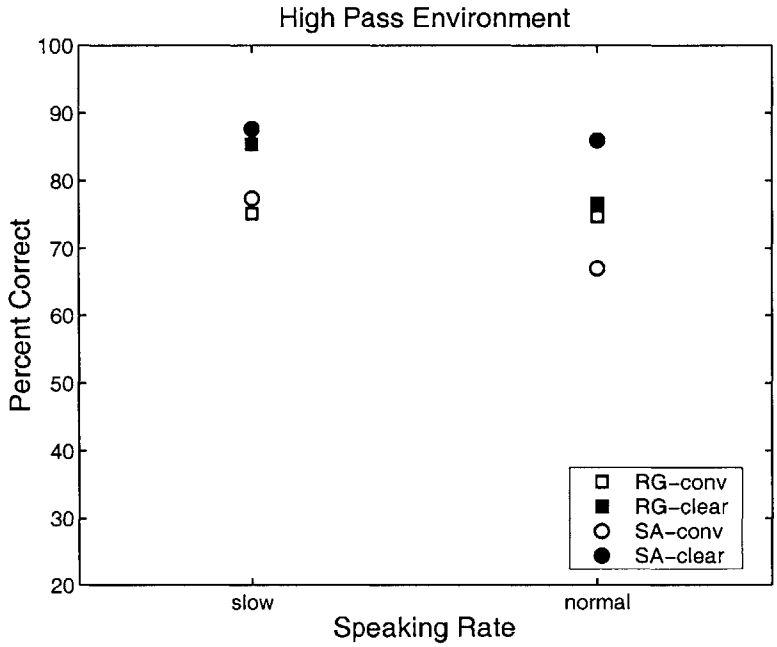

Figure 4-3: Intelligibility data versus rate in a high pass environment. Squares indicate RG and circles indicate SA. Open symbols are data for conversational speech, and filled symbols are data for clear speech.

speech-shaped noise was added to the filtered speech at roughly 35dB below the peak speech level in order to ensure that any remaining out of band speech energy did not affect intelligibility.

Five additional normal-hearing listeners were employed to evaluate the intelligibility of this processing. Again, one hundred nonsense sentences from each talker in conv/normal and conv/slow modes and fifty sentences from each talker in clear/normal and clear/slow modes were presented. The percent-correct key word intelligibility scores for each talker, averaged across listener, are presented in Figure 4-3. At both slow and normal rates, SA achieved an intelligibility benefit in this environment by speaking clearly, whereas RG only achieved the improvement at a slow speaking rate.

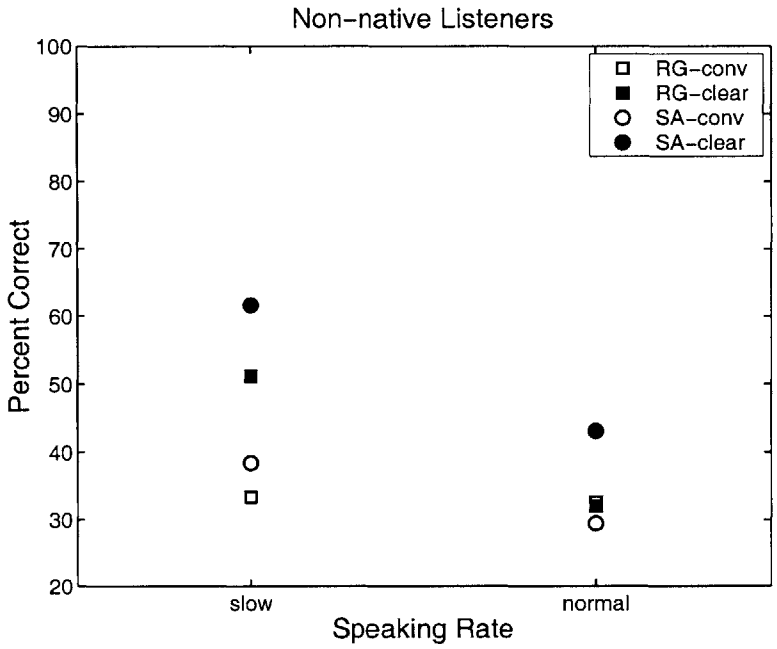

Figure 4-4: Intelligibility data versus rate in for non-native listeners. Squares indicate RG and circles indicate SA. Open symbols are data for conversational speech, and filled symbols are data for clear speech.

Non-native Listening Environment

In the final difficult listening environment, three normal-hearing listeners who learned English as a second language were employed to evaluate the intelligibility of the sentences in additive speech-shaped noise[39] at an SNR of 0dB. In this environment, 50 nonsense sentences from each talker were presented in conv/normal, clear/normal, conv/slow, and clear/slow modes. The percent-correct key word intelligibility scores for each talker, averaged across listener, are shown in Figure 4-4. Similar to the reverberant and high-pass environments, SA achieved an intelligibility improvement by speaking clearly at both rates, while RG only showed such an improvement at the slow rate.

Types of Clear Speech

The results showed that SA's clear/normal speech was more intelligible than his conversational speech in three of the conditions (reverberation: 19 percentage points; high-pass: 19 percentage points; non-native listeners: 14 percentage points). The only condition in which his clear/normal speech did not provide an intelligibility advantage was for the low-pass filtering condition. In this condition, his conversational speech was so highly intelligible (87%) that a ceiling effect could account for his failure to improve his intelligibility. RG's clear/normal speech, however, was significantly more intelligible than her conversational speech in only one of the four conditions (low-pass: 15 percentage points). Thus, SA's strategy for producing clear/normal speech appears to be more robust to other degradations than RG's strategy, suggesting that signal processing schemes aimed at enhancing intelligibility should be modeled after SA's strategy, in order to maximize benefit to the listener. In addition, the results suggest that different types of clear speech may exist, depending on the talker and the environment.

If this is the case, it would then be reasonable to expect that additional types of clear speech could be elicited from talkers by varying either the type of distortion employed or the constraints imposed on speaking rate. Furthermore, the intelligibility of each type of clear speech may then vary for different distortions. A similar effect has been shown for different styles of conversational speech. Dix[9] has shown that one style of conversational speech can be significantly more intelligible than another style for a reverberation distortion, even though the two styles of speech have comparable intelligibility under low-pass conditions.

The question of whether and how many types of clear speech exist has yet to be explored in clear speech research. This work is not intended to characterize all types of clear speech but rather, to focus on determining the properties of the specific type(s) of clear/normal speech that have been produced by RG and SA and shown to be resistant to an additive noise distortion.

4.3 Summary and Conclusions

Multivariate statistical analyses revealed no linear combinations of multiple acoustical variables that could be responsible for the high intelligibility of clear speech. Thus, the acoustical differences identified between clear/normal and conv/normal speech in Chapter 3, summarized in Table 4.3, are the most likely indicators of characteristics necessary for highly intelligible speech. However, some of these characteristics may be more likely than others to account for significant portions of the intelligibility advantage, since the intelligibility tests in other degraded environments, summarized in Table 4.4 suggest that the talkers may have used somewhat different strategies for producing clear/normal speech and that SA's strategy may have been more successful. In particular, acoustical differences exhibited by both talkers or by SA, whose clear/normal speech was more robust to various degradations, may be more likely linked to significant intelligibility gains.

Table 4.3: Summary of acoustical properties of clear/normal speech relative to conv/normal speech, identified in Chapter 3.

Property	Difference	Talker SA	Talker RG
Long-term spectra	more energy above 1kHz	Yes	Yes
Short-term vowel spectra	increased energy near 2nd and 3rd formants	Yes	Yes
Temporal envelopes	increased modulation index for frequencies <3–4Hz	Yes, in 4 of 7 octave bands	No (only in 1 octave band)
F0	greater average, range	Yes	No
Word-initial stops	increased VOT	Yes	No
Word-final stops	bursts released more often	No	Yes

Of the six acoustic differences identified in Chapter 3, two were exhibited by both talkers, three were exhibited by SA only, and one was exhibited by RG only. The two differences exhibited by both talkers (relative energy changes in short-term and long-term spectra) both represent the same acoustic property, increased energy near second and third formants. The RG-only difference (releasing more stop bursts in clear/normal speech) does not have a high potential for enhancing intelligibility in a large

Table 4.4: This table summarizes the results of intelligibility tests presented in this Chapter, indicating whether each talker's clear/normal speech was more intelligible than his/her conv/normal speech for the specified environment.

Environment	Talker SA	Talker RG
Reverberation	Yes	No
Low-Pass	No (possible ceiling effect)	Yes
High-Pass	Yes	No
Non-native	Yes	No

number of settings, because RG's clear/normal speech does not have an intelligibility advantage in a wide variety of environments. Of the SA-only differences, lengthening VOT of stop consonants is not likely responsible for substantial improvements in intelligibility, since so few word-initial stop consonants are in the database. The other three properties of clear/normal speech, exhibited either by SA only or by both talkers, however, seem likely to contribute to its high intelligibility. Therefore, signal processing schemes that further investigate these properties (vowel formant energy, F0 average and range, and low-frequency modulations) are explored in the following chapters.

Chapter 5

Signal Transformations Aimed at Intelligibility Enhancement

Signal transformations were developed and applied to conv/normal speech. These transformations altered the following three properties of conv/normal speech: A) vowel formant energy was increased by raising formant amplitudes, B) the fundamental frequency (F0) was modified to increase the average and expand the range of F0 values, and C) low-frequency modulations (<3–4Hz) of the intensity envelopes were enhanced in several octave bands.

This chapter discusses the details of the processing schemes and their effects on the acoustics of the speech waveform, both individually and in combination. Intelligibility tests were also conducted to determine if the processing schemes were beneficial to intelligibility singly or in combination. Those results are discussed in Chapter 6.

5.1 Formant Frequencies

Because there was relatively more energy near the second and third formants in clear speech relative to conversational speech (see Chapter 3), it seems likely that a signal transformation which increases power to these formants may be helpful in improving intelligibility. Such a transformation was implemented by modifying the magnitude of the short-time Fourier transform (STFT) and then using the Griffin-Lim[17] algorithm

to estimate a signal from its modified STFT magnitude.

The first step of processing was to calculate the STFT magnitude, or spectrogram, using an 8ms Hanning window with 6ms of overlap. Next, the formant frequencies were measured at 10ms intervals for voiced portions of the speech signal using the formant tracking program provided in the *ESPS/waves+* software package. For each 10ms interval where voicing was present, the spectrogram magnitude was multiplied by a modified Hanning window, $w[F]$, whose endpoints in frequency, F_{start} and F_{end} are represented by the following:

$$F_{start} = F_2 - min(2BW_2, \frac{F_1 + F_2}{2}) \tag{5.1}$$

$$F_{end} = F_3 + min(2BW_3, \frac{F_2 + F_3}{2}) \tag{5.2}$$

where F_2, F_3, BW_2, and BW_3 are the second and third formants and their bandwidths, respectively. A Hanning window spanning this frequency range, $h[F]$, was modified according to the following formula:

$$w[F] = Ah[F] + 1 \tag{5.3}$$

where A was a scale factor, that could be used to determine the amount of amplification. For the processing in this study, 2 was found to be a good value for A. Finally, the Griffin-Lim[17] iterative algorithm was used to estimate the signal from the enhanced spectrogram. The sentences were then normalized for long-term RMS value, and the effect on the long-term and short-term spectra was examined.

5.1.1 Effect on Long-term RMS Spectra

The long-term spectra of conv/normal, clear/normal, and formant processed/normal speech, normalized for long-term RMS level, were computed as in Section 3.2.3. The spectral differences of clear and processed modes relative to conversational speech are shown for each talker in Figures 5-1 and 5-2, demonstrating that the processing had the desired effect on the long-term spectrum. At frequencies above 1kHz, the pro-

cessed speech exhibits roughly the same increase in energy, relative to conv/normal speech, as clear/normal speech does.

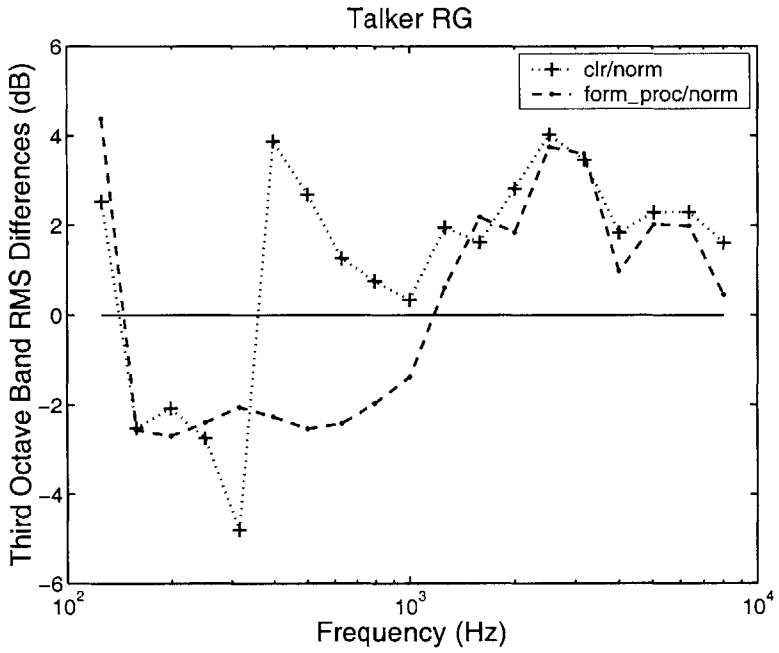

Figure 5-1: Third-octave band RMS spectral differences of RG's clear/normal and (formant) processed/normal modes relative to conv/normal speech.

5.1.2 Effect on Short-term RMS Spectra

The short-term spectra of conv/normal, clear/normal, and formant processed/normal speech, normalized for segment RMS level, were computed as in Section 3.4.3. Typical results for consonants are shown in Figures 5-3 and 5-4. Since neither talker exhibited much spectral change between conv/normal and clear/normal styles for consonants, the spectra obtained for these processed/normal consonants is appropriate.

Typical vowel spectra for each talker in are shown in Figures 5-5 and 5-6. As in Chapter 3, the vowels displayed span the range of F1 (correlated with vowel height) and F2 (correlated with vowel fronting) and thus are fairly representative of the entire vowel space. These plots show that the processing increases energy near the second and third formants, a result that holds for nearly all other vowels as well (see

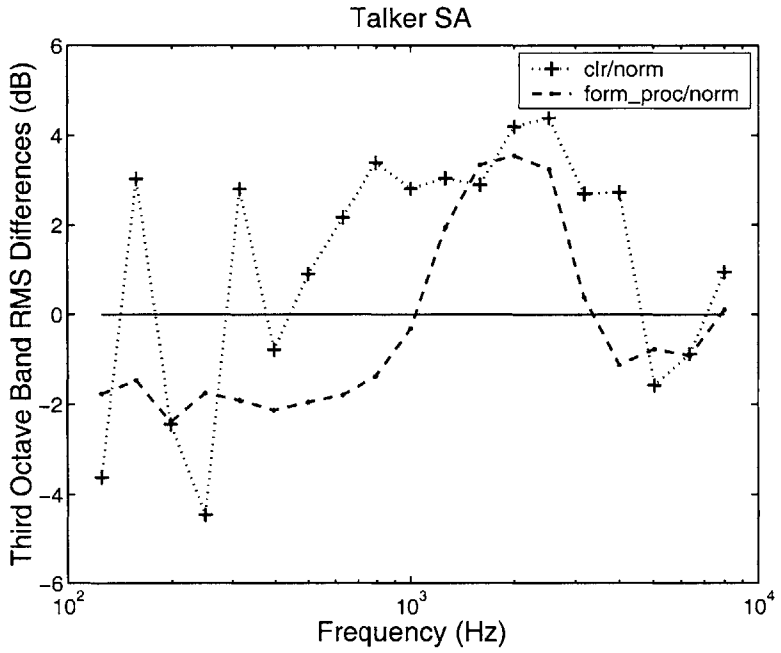

Figure 5-2: Third-octave band RMS spectral differences of SA's clear/normal and (formant) processed/normal modes relative to conv/normal speech.

Figures B-1 through B-8 in Appendix B).

5.2 Fundamental Frequency

Using a method based on linear prediction coefficients (LPC), conversational speech was processed to widen the fundamental frequency (F0) range as well as to increase the average F0 value. Although this pattern of change in F0 behavior was observed for clear speech relative to conversational speech only for male talkers (see Chapter 3), the processing was applied to both male and female talkers.

The conversational speech to be processed was first analyzed with a suite of LPC programs available in the *ESPS/waves+* toolkit. With these programs, the pitch epochs (period of vocal fold closure during voiced portions of the speech signal) were determined and pitch-synchronous LPC analysis was performed. In addition, fundamental frequency (F0) values were extracted at 10ms intervals from voiced portions of the speech signal. These F0 values were then modified so that the average value

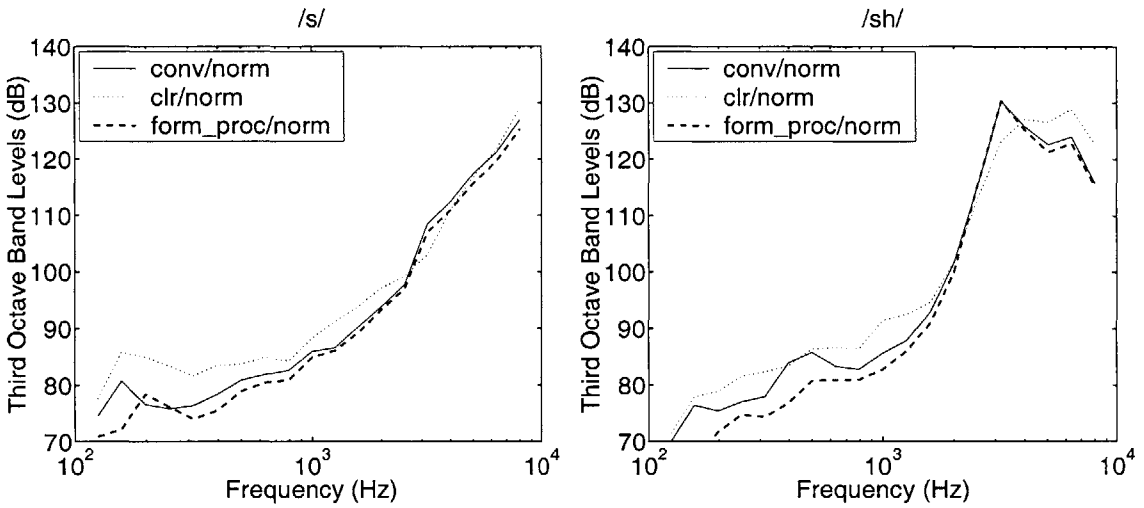

Figure 5-3: Third-octave average spectra of /s/ and /sh/ in conv/normal, clear/normal, and processed/normal modes for RG. Similar results (no significant difference between modes) were obtained for all consonants.

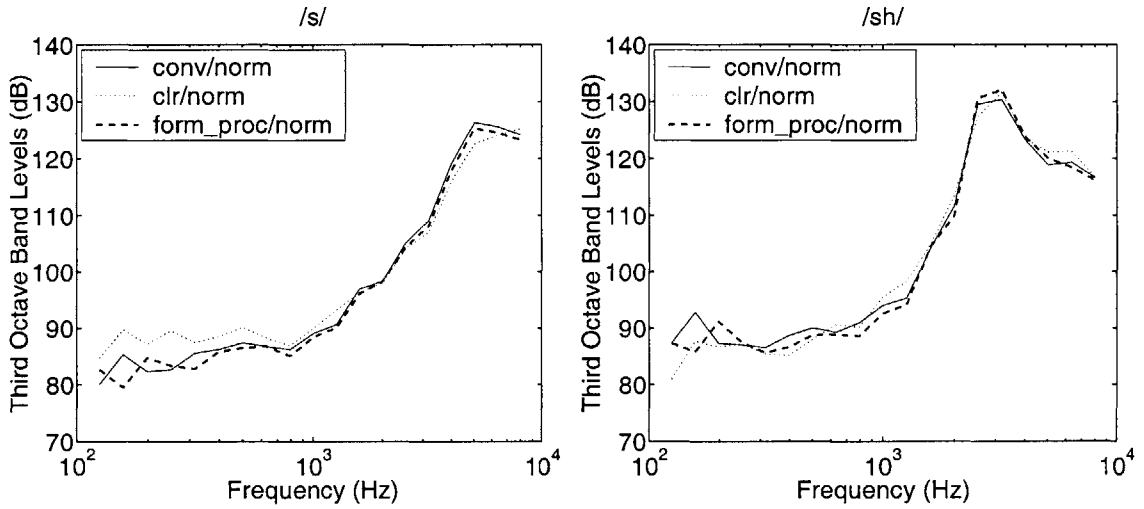

Figure 5-4: Third-octave average spectra of /s/ and /sh/ in conv/normal, clear/normal, and processed/normal modes for SA. Similar results (no significant difference between modes) were obtained for all consonants.

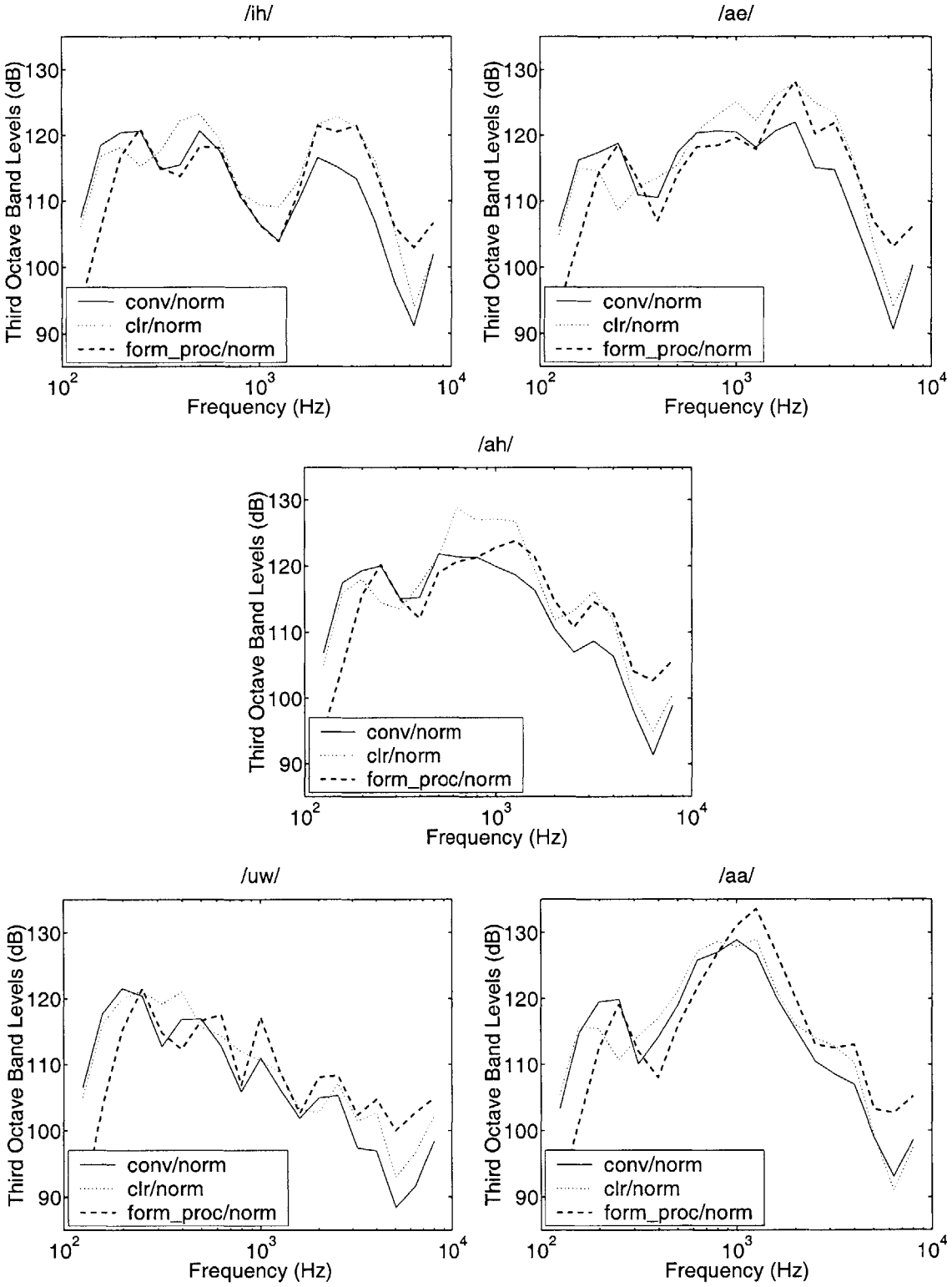

Figure 5-5: Third-octave average spectra for five vowels (spanning F1-F2 space) in conv/normal, clear/normal, and processed/normal modes for RG. Similar results were obtained for most vowels. Spectra for remaining vowels can be found in Appendix B.

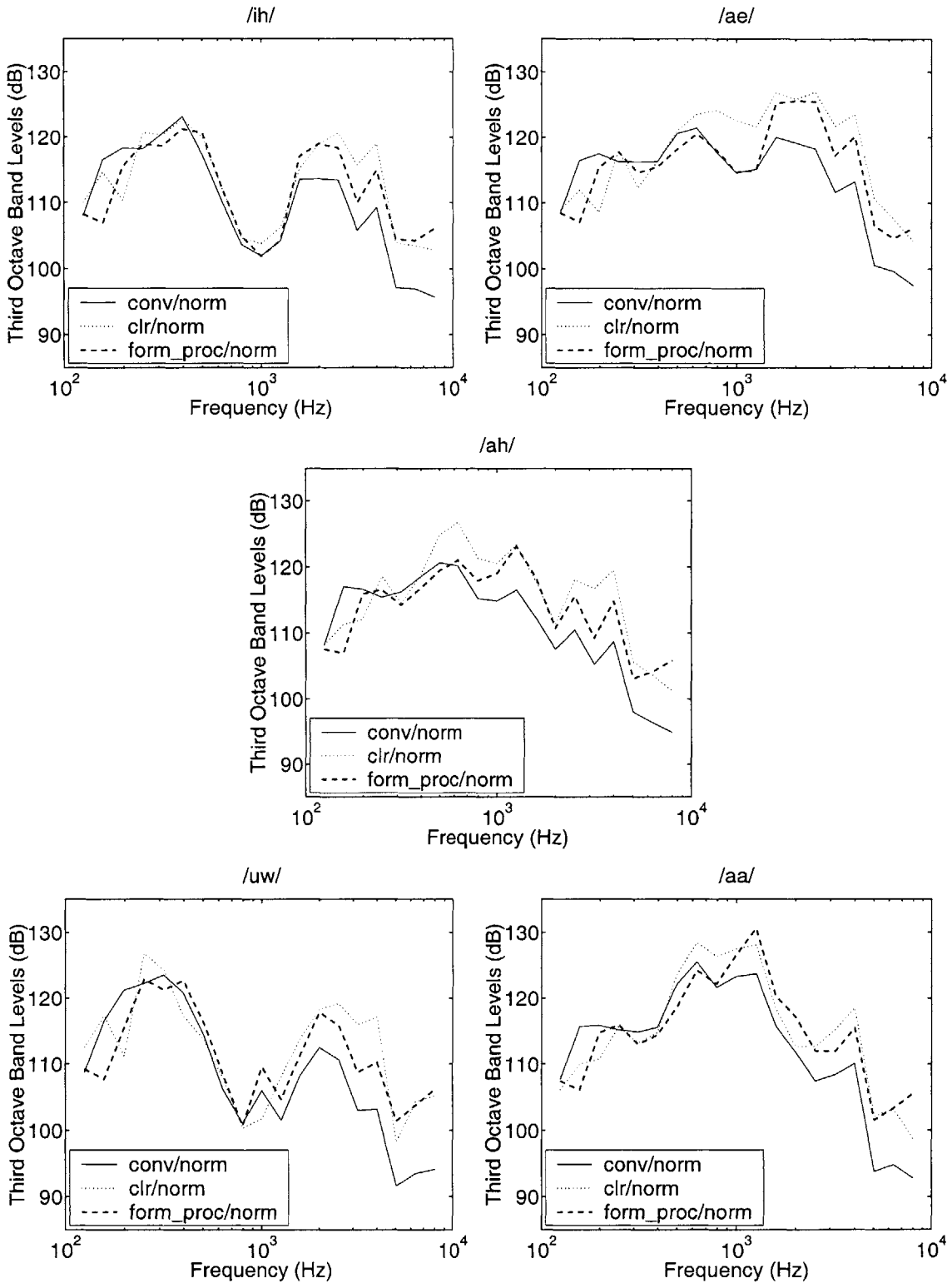

Figure 5-6: Third-octave average spectra for five vowels (spanning F1-F2 space) in conv/normal, clear/normal, and processed/normal modes for RG. Similar results were obtained for most vowels. Spectra for remaining vowels can be found in Appendix B.

for each list of 50 sentences was increased by I Hz, and the standard deviation for each list was expanded by a factor of g, according to the following formula:

$$F0\_enh = g(F0\_orig - F0\_avg) + F0\_avg + I.$$

LPC synthesis was then performed using the modified pitch track and LPC coefficients for each sentence to create an enhanced version of the sentence. Finally, the enhanced sentences were normalized for long-term RMS level.

The values for I and g were initially set to 41 Hz and 1.9, respectively, which were the average values observed for those sentence lists that showed a significant change in F0 with speaking mode in Chapter 3. These values were used for processing the speech of male talkers. However, it was determined through informal listening tests that these values were too large for female talkers and caused the speech to sound very unnatural. Therefore, values roughly 1/2 as large (20 Hz for I and 1.45 for g) were used for processing the speech of female talkers.

In order to evaluate the effect of this processing on the pitch track, fundamental frequency (F0) values were extracted at 10ms intervals from voiced portions of the processed speech. A histogram of the F0 values obtained for the two talkers analyzed in Chapter 3, with reference values for natural speech, is shown in Figure 5-7. Appendix B also contains a summary of means and standard deviations for each talker and speaking style (see Table B.2). The processing had the intended effect on the pitch tracks for each gender. For the male talker (SA), the processed/normal F0 values very closely resemble the F0 values for clear/normal speech and are similar to the clear/slow values. For the female talker (RG), the processed/normal F0 values have a greater range and higher average value than for both clear/normal and clear/slow styles of speech.

As in the pitch analysis of Chapter 3, a crude estimate of sentence level intonation contours was calculated by measuring F0 at two points in each sentence: at its maximum value and 50ms before the end of the sentence. These values were averaged over 50 sentences for each speaking style. The results for both talkers are plotted

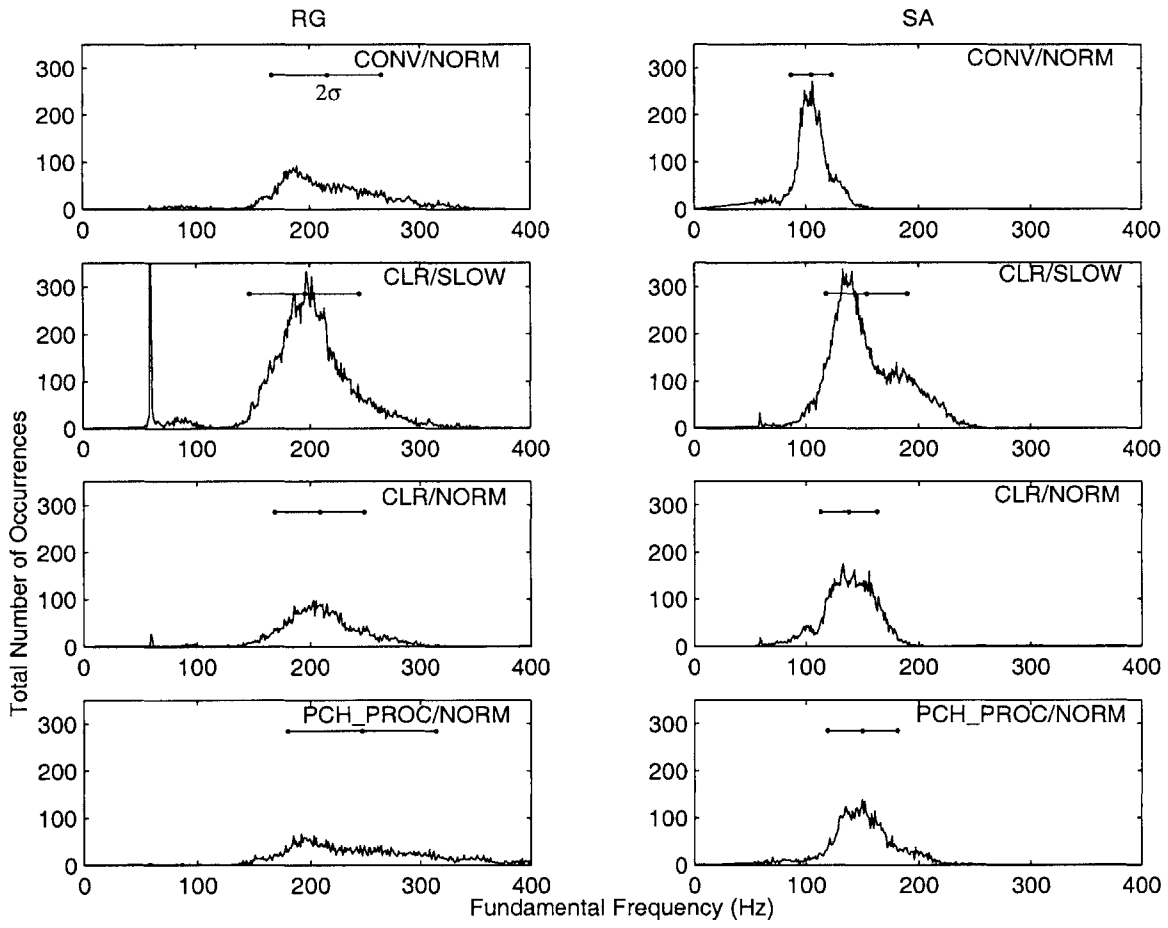

Figure 5-7: Fundamental frequency distributions. Each row shows distributions for different speaking modes; columns give results for each talker.

in Figure 5-8. Again, the processing had the expected effect for each gender. For SA, the male talker, the processed/normal intonation contour falls between the contours for clear/normal and clear/slow, whereas for RG, the female talker, the entire processed/normal contour falls above the contour for all of the other speaking styles.

5.3 Temporal Envelope

While the increase in modulation depth for frequencies less than 3–4Hz observed in Chapter 3 is not as great in clear/normal speech as in clear/slow speech, the trend is still substantial for talker SA in four of the seven octave bands (250Hz,

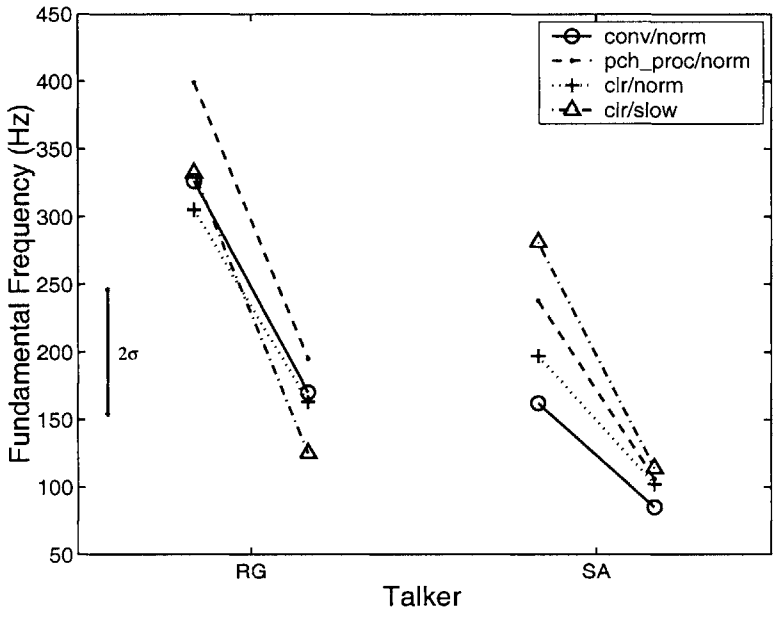

Figure 5-8: The maximum F0 value attained in a sentence followed by the value 50ms before the end of the last word in the same sentence, averaged over 50 sentences (a rough approximation of the F0 contour) for each speaking style.

500Hz, 1000Hz, and 2000Hz). Although the increase was observed in all octave bands for clear/slow speech, it was assumed that the enhancement of slowly varying modulations in the 125Hz, 4000Hz, and 8000Hz bands was likely a result of the reduction in speaking rate associated with clear/slow speech. Therefore, the third signal enhancement transformation was designed to increase the modulation depth for frequencies less than 3–4Hz in the intensity envelope of the 250Hz, 500Hz, 1000Hz, and 2000Hz octave bands.

A block diagram of the analysis-synthesis scheme used for this transformation is depicted in Figure 5-9. First, in the analysis stage, the downsampled intensity envelope and envelope spectrum for a list of 50 sentences were computed as in Section 3.2.4. The transfer characteristics of the eight octave-band filters used in this stage of processing can be found in Figure 5-10. The filters were designed so that the overall response of the combined filters, shown in Figure 5-11, was roughly $\pm 2dB$ over the entire range of speech frequencies.

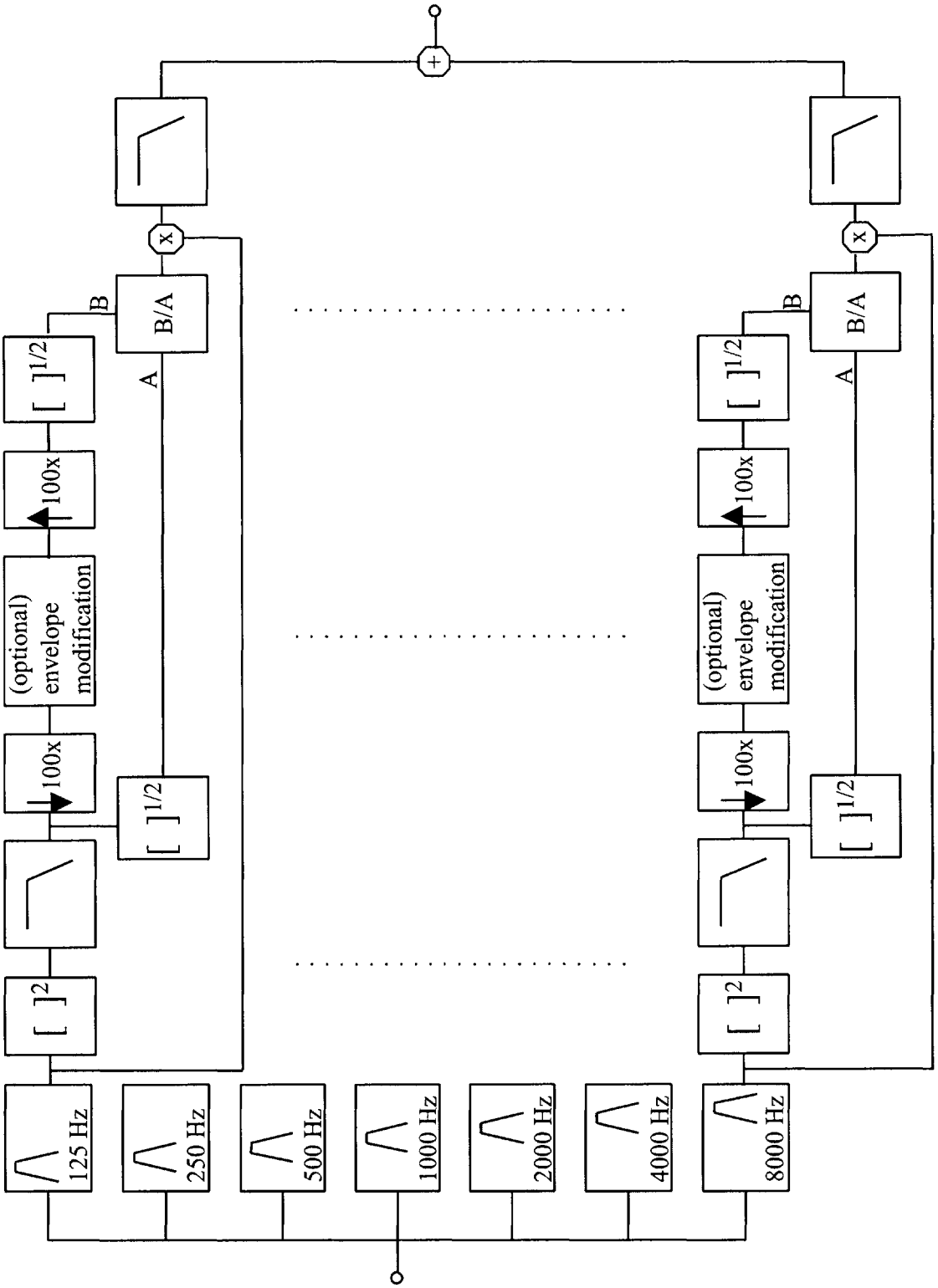

Figure 5-9: Block diagram of temporal envelope processing scheme.

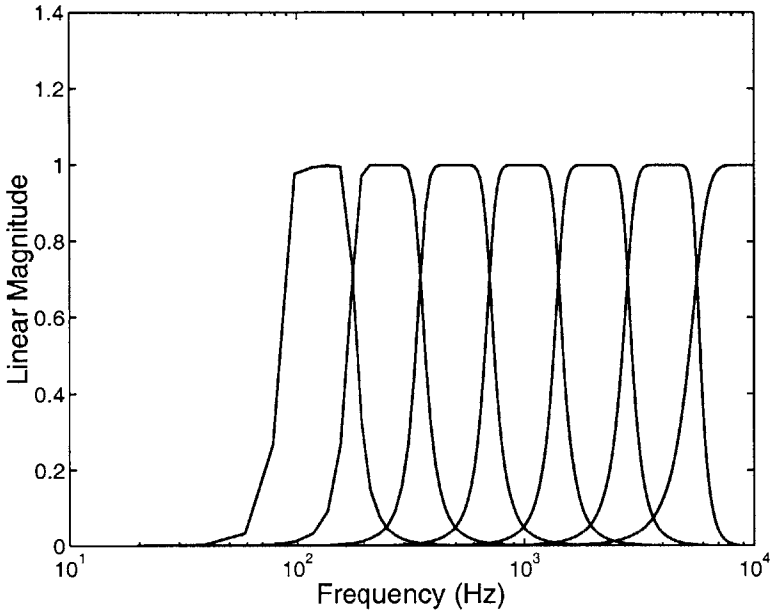

Figure 5-10: Transfer characteristics of eight octave-band filters used in processing temporal envelopes.

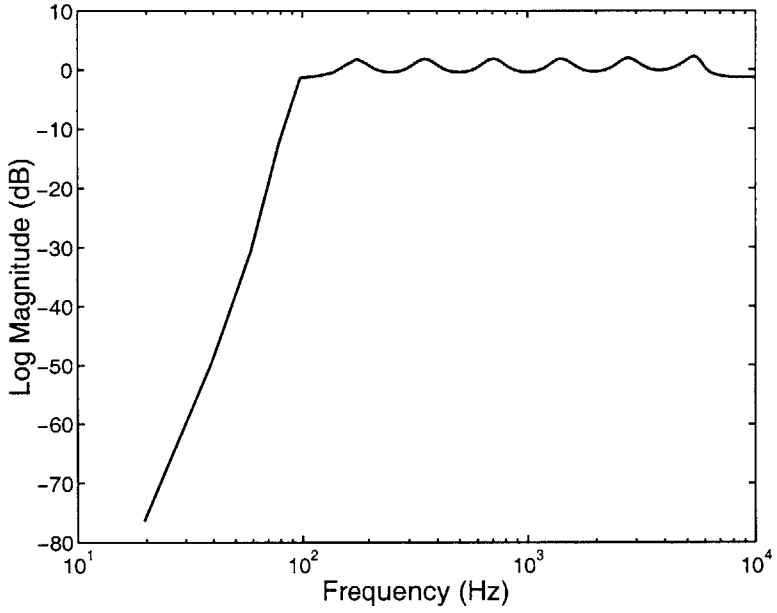

Figure 5-11: Overall filterbank transfer characteristics.

In the second stage of processing, the envelope of each sentence in the list was processed separately. For the octave bands in which modification was desired, the original envelope of each sentence was modified by a 200-point FIR filter designed to amplify frequencies between 0.5 Hz and 4 Hz. The resulting modified envelope was also adjusted to have the same average intensity as the original sentence envelope, and then any negative values of the adjusted envelope were set to zero. If resetting the negative portions of the envelope to zero affected the average intensity substantially, the intensity adjustment procedure was repeated until the average intensity of the modified envelope was within 0.5% of the average intensity of the original envelope. The modified envelope and original envelope were then upsampled to the original sampling rate of the signal in order to prepare for the final synthesis stage of processing.

Although it was desirable to work with intensity envelopes in the first two stages of processing so that the desired intensity envelope spectra could be achieved in the modified signal, the *amplitude* envelope was necessary for synthesis. Therefore, during the final processing stage, the time-varying ratio of the amplitude envelopes was calculated by comparing the square-root of the modified intensity envelope with the square-root of the original intensity envelope. The original octave band signals were then transformed by multiplying the original signal in each octave band (with fine structure) by the corresponding time-varying amplitude ratio for that band. In order to ensure that no energy outside the octave band was inadvertently amplified, the result was also low-pass filtered by a 4th-order Butterworth filter with cutoff frequency corresponding to the upper cutoff frequency for the octave band. The processed version of the signal was then obtained by summing the signals in each octave band. Lastly, the processed sentences were normalized for long-term RMS level.

After synthesis was complete, it was determined through informal listening tests that these modifications caused the speech of female talkers to sound more unnatural than the male talker. A likely explanation for this problem is that the fundamental frequency of the female talkers tends to fall in the second octave band, and amplifying slowly varying modulations of voicing is not likely to occur in natural speech unless

the talker slows down. This is supported in the data, since an increase in modulation index is not exhibited in the 250Hz band for clear/normal speech for the female talker, RG, although it is present in the 500Hz band (see Section 3.2.4). As a result, the signal transformation procedure was specified to modify only the 500Hz, 1000Hz, and 2000Hz bands for female talkers.

In order to evaluate the effect of the processing on the intensity envelopes, the envelope spectra of the processed speech was computed as in Section 3.2.4. The spectra of the octave-band intensity envelopes for each talker's processed speech are shown in Figures 5-12 and 5-13. Spectra for the unprocessed speech (conv/normal), as well as the two naturally clear conditions (clear/normal and clear/slow) are provided in these figures for comparison purposes. From these figures, it can be seen that the processing had the desired effect, with the envelope spectra of processed/normal speech falling between that of clear/normal and clear/slow for frequencies less than 3–4Hz in the specified octave bands (500Hz, 1000Hz, and 2000Hz for both talkers as well as 250Hz for SA).

5.4 Combination of Schemes

To attempt to approximate all of the acoustic properties identified in clear/normal speech simultaneously, a combination transformation was created by applying each of the above signal processing schemes in succession. Since the individual transformations were not linear, the order in which the transformations were applied affected the final outcome. Through informal listening tests, it was determined that the combination ordering with the least amount of artifacts consisted of formant processing, followed by fundamental frequency processing, followed by temporal envelope processing.

In order to evaluate whether the altered acoustic characteristics (long-term spectrum, fundamental frequency, and temporal envelope modulation) remained near their intended levels at the conclusion of the combination processing, each was measured as described in the sections above. The results are shown in Figures 5-14 through 5-18.

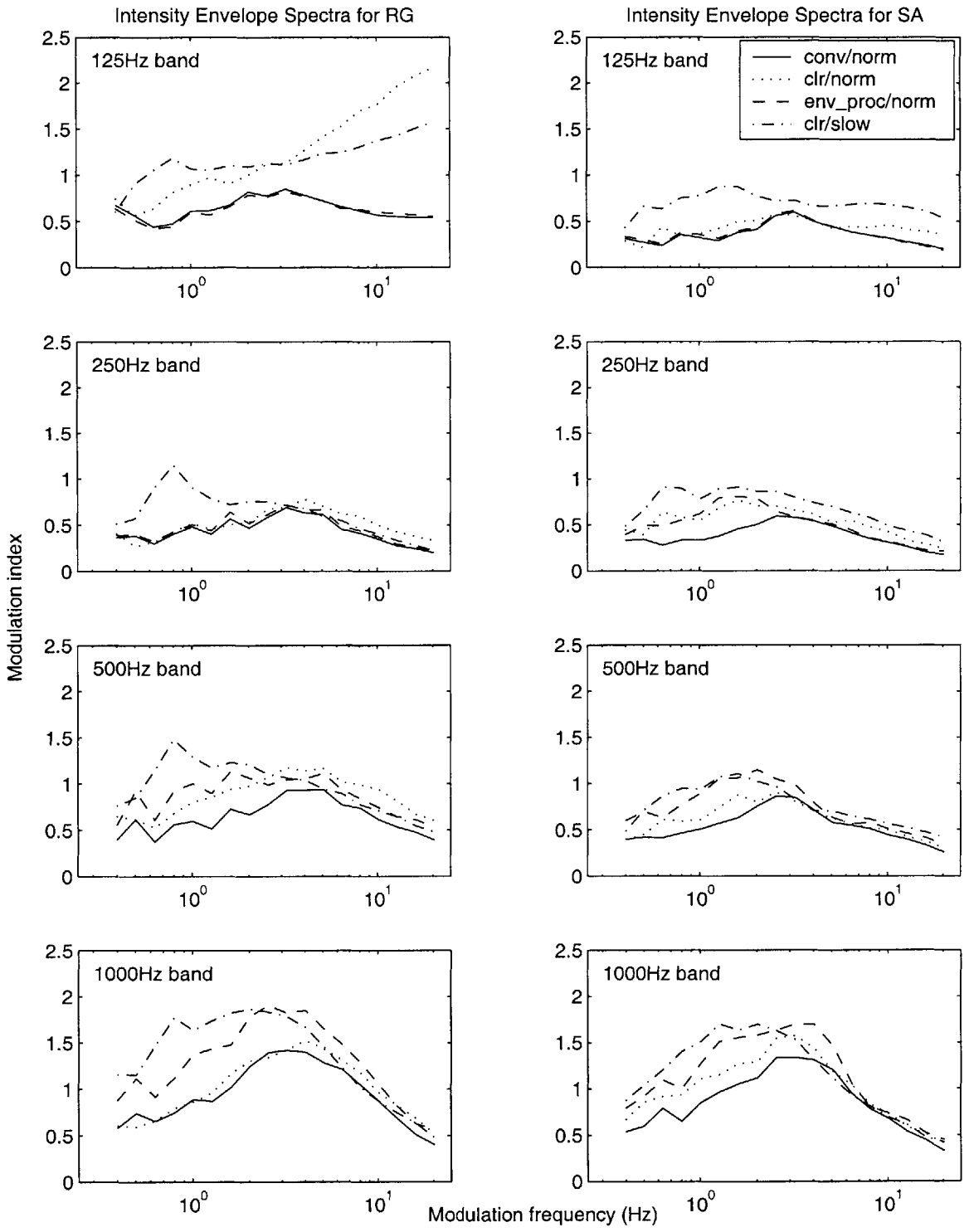

Figure 5-12: Spectra of intensity envelopes for Talkers RG and SA in lower four octave bands.

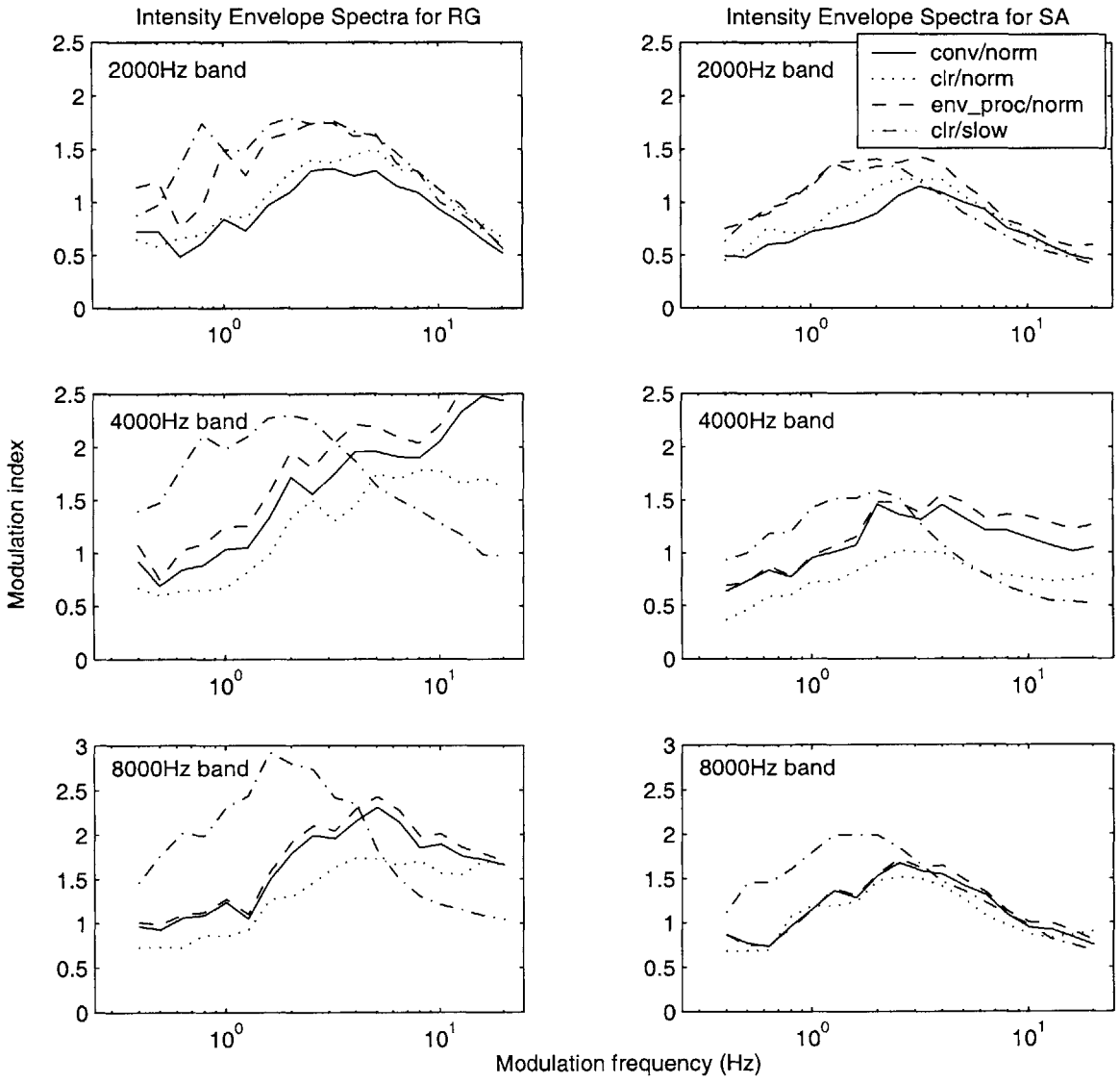

Figure 5-13: Spectra of intensity envelopes for Talkers RG and SA in upper three octave bands.

Long-term spectral differences of clear and processed modes relative to conversational speech are shown for each talker in Figures 5-14 and 5-15. These figures show that the relative increase in frequencies above 1kHz remained in effect after the combination processing was applied. A histogram of pitch values for each mode is shown in Figure 5-16, demonstrating that the intended effect on F0 distribution was present in the (combination) processed speech. Thus, both the long-term spectrum and the fundamental frequency of the (combination) processed speech very closely matched the desired values.

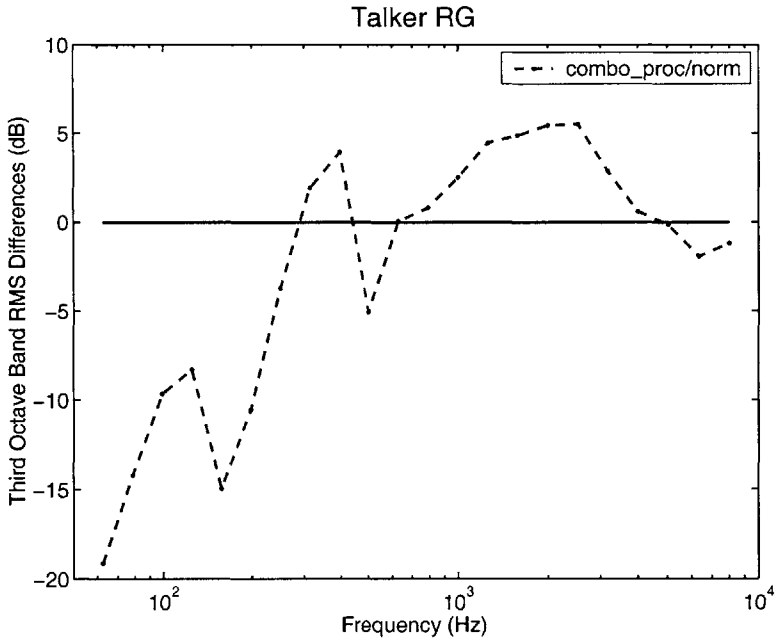

Figure 5-14: Third-octave band RMS spectral differences of RG's clear/normal and (combination) processed/normal modes relative to conv/normal speech.

The effect of the combination processing on temporal envelope modulations, however, was somewhat different than desired. Figures 5-17 and 5-18 depict the spectra of the octave-band temporal intensity envelopes for each talker's conv/normal and (combination) processed/normal speech. From these figures, it can be seen that in addition to the intended effect of increasing the modulation depth for frequencies less than 3–4Hz for most octave bands (500Hz, 1000Hz, and 2000Hz octave bands for both talkers as well as in the 250Hz band for SA), the modulation depth of frequencies

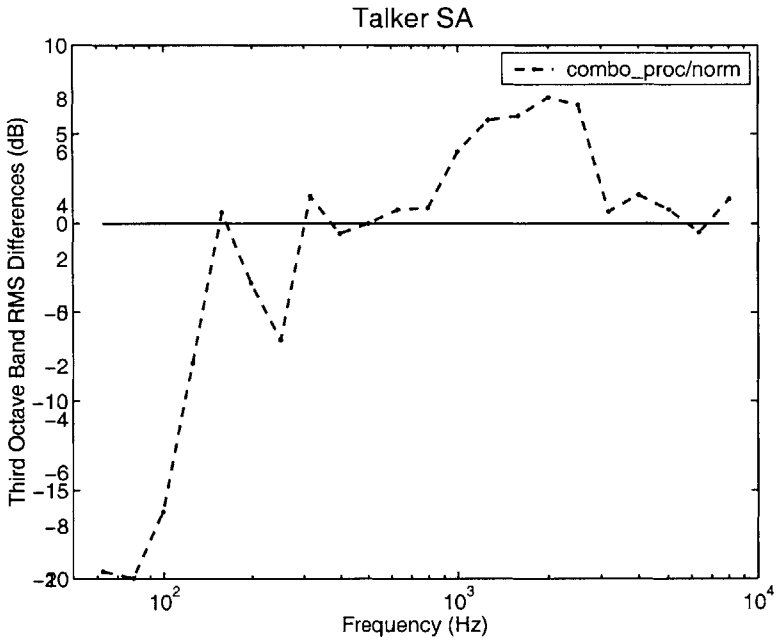

Figure 5-15: Third-octave band RMS spectral differences of SA's clear/normal and (combination) processed/normal modes relative to conv/normal speech.

above 4Hz was also affected in many of the octave bands. In most bands for SA and in the 500Hz and 2000Hz bands for RG, the amount of increase for the high modulation frequencies was generally less than the amount of increase seen for modulations less than 4Hz, thus preserving a relative increase in depth of low frequency modulations. Nonetheless, the effect of the unintended change in high frequency modulations on intelligibility was unknown. This question was investigated theoretically by applying the Speech Transmission Index. The results of this investigation are presented in Chapter 6.

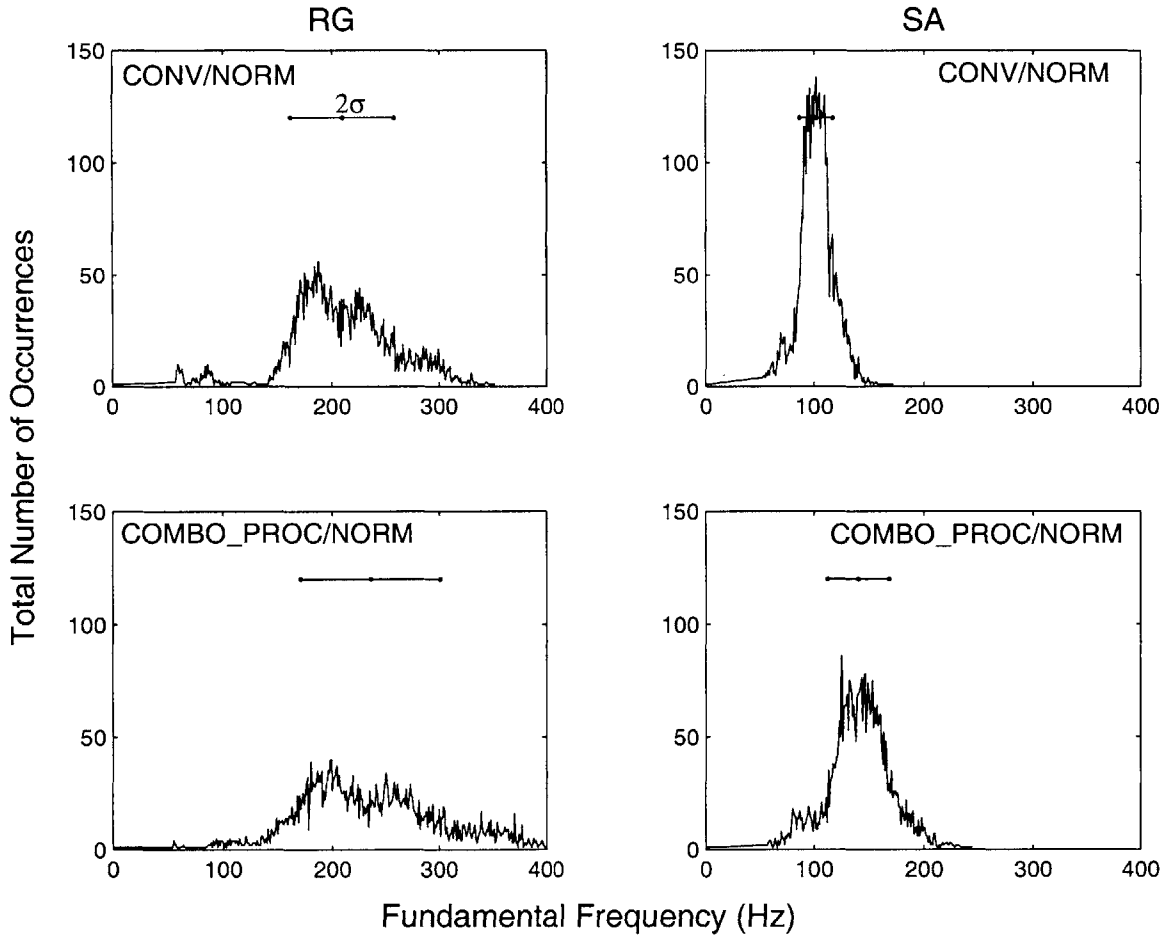

Figure 5-16: Fundamental frequency distributions for the speech of SA and RG after applying the signal transformations in combination. Each row shows distributions for different speaking modes; columns give results for each talker.

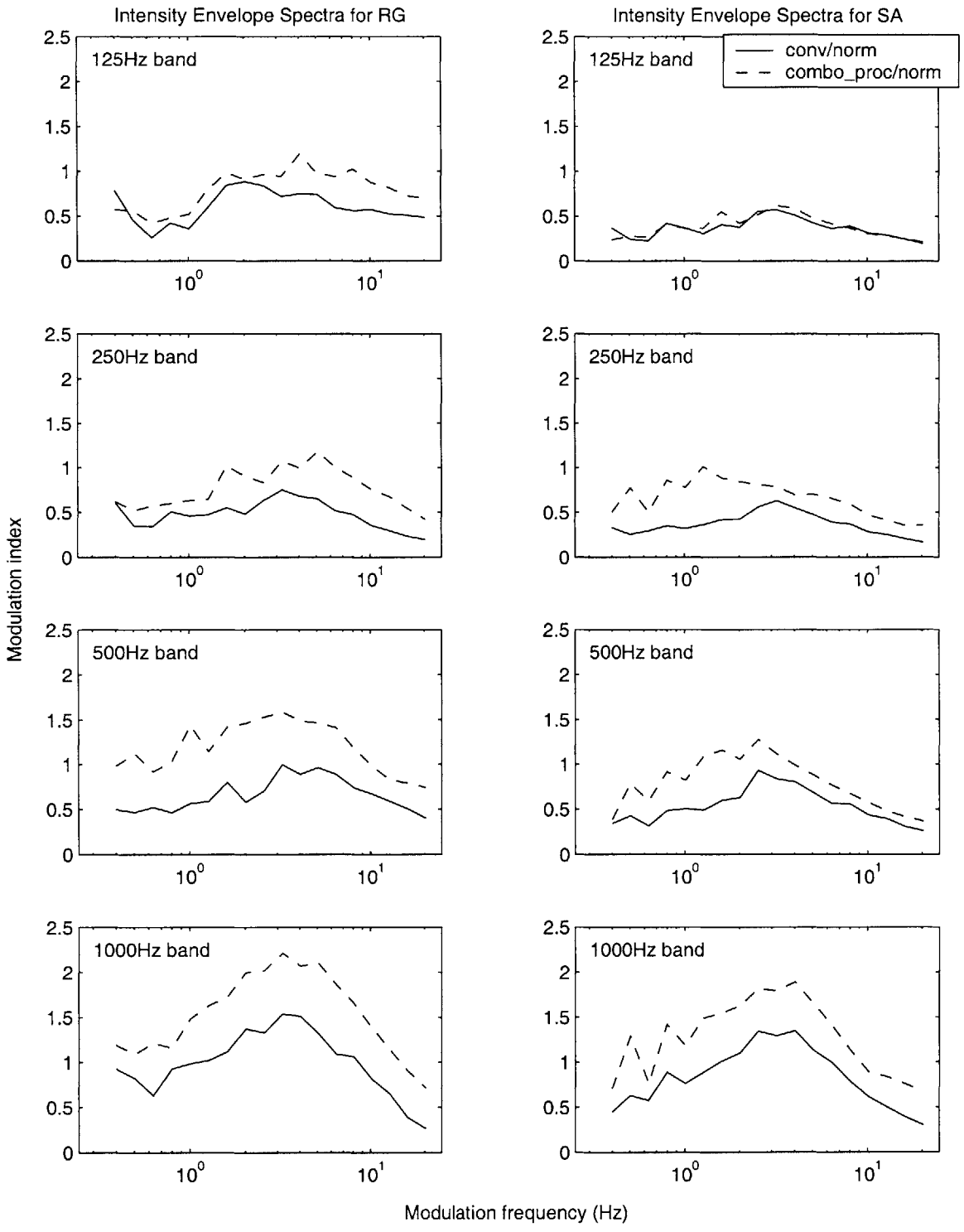

Figure 5-17: Spectra of intensity envelopes, before and after applying the three signal transformations in combination, for Talkers RG and SA in lower four octave bands.

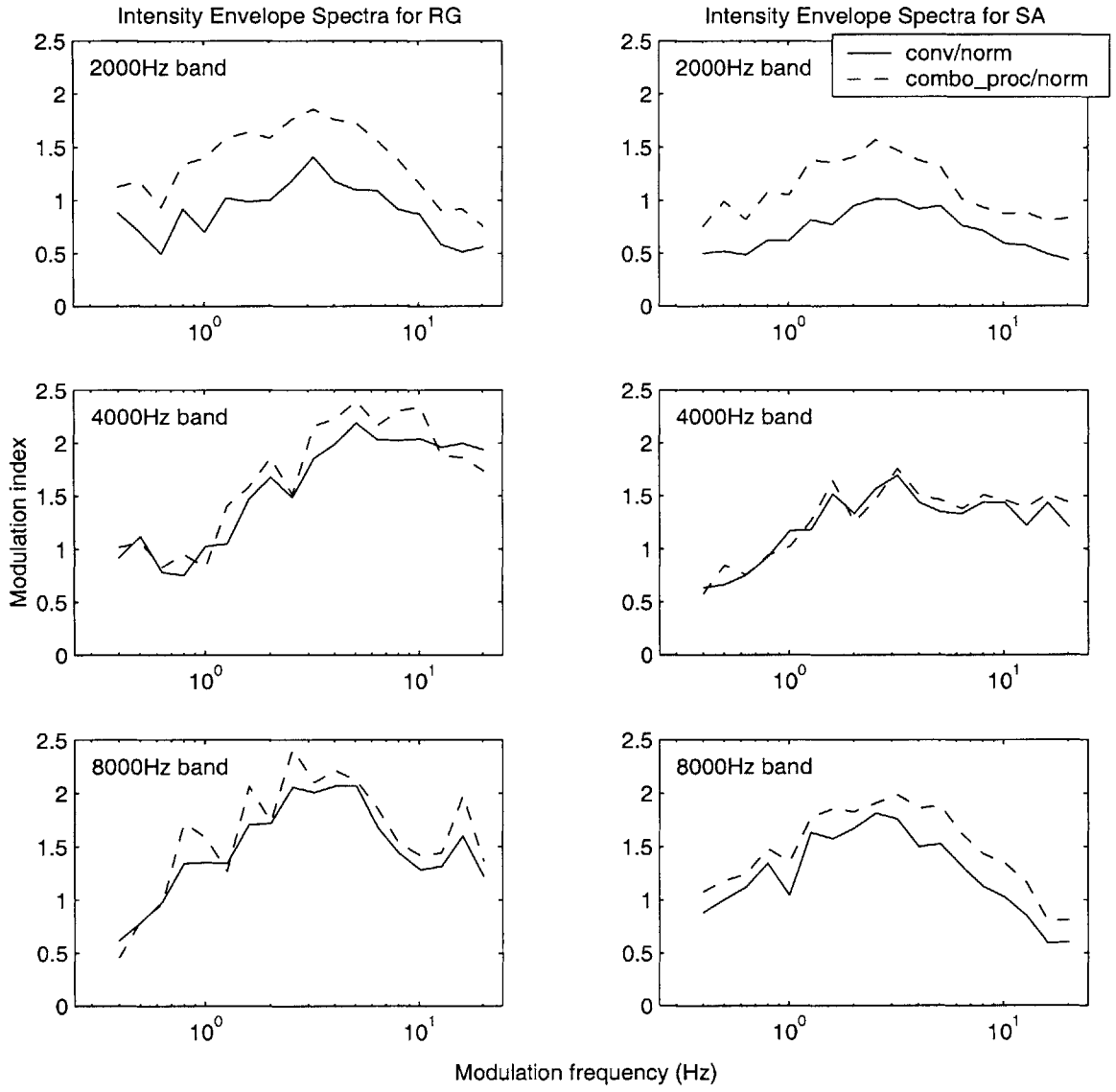

Figure 5-18: Spectra of intensity envelopes, before and after applying the three signal transformations in combination, for Talkers RG and SA in upper three octave bands.

5.5 Summary

Based on the data in Chapter 3, it seems possible that one or more of the transformations described in this chapter, alone or in combination, could produce an increase in intelligibility. Each individual processing scheme appeared to have roughly the desired effect on the appropriate acoustic characteristic of the signal. The combination of all three processing schemes resulted in speech that exhibited each of the three desired acoustic properties of highly intelligible speech but also had increased high frequency (>4Hz) modulations. The effect of this unintended alteration on intelligibility is unknown. However, the potential detriment of this acoustic characteristic and any other possible signal processing artifacts was assessed with the Speech Transmission Index. The theoretical predictions of speech intelligibility obtained from the STI and the results of actual intelligibility tests for normal hearing and hearing impaired subjects are reported in Chapter 6.

Chapter 6

Evaluation of Intelligibility Enhancement Schemes

To assess the effectiveness of the processing schemes discussed in Chapter 5, the intelligibility of several speaking styles was predicted using the speech-based STI measure and the actual intelligibility of these speaking styles was measured in two separate intelligibility experiments. In the first experiment, the speech stimuli were presented to normal hearing listeners in the presence of wideband noise, and in the second, the speech stimuli were presented to impaired listeners in a quiet background.

6.1 Speech Stimuli

The speech stimuli used for these experiments consisted of a subset of speech recorded for Krause's earlier study of clear speech naturally elicited at normal speaking rates[28] (nonsense sentences from the Picheny corpus[44]). In Krause's study, four of five talkers were able to produce clear speech at normal rates or faster. The conv/normal, clear/normal, and conv/slow speech of those talkers was used for these intelligibility experiments. In one case (Talker EK), clear/quick speech was used rather than clear/normal, since Krause's study showed higher scores at the quick rate than at the normal rate. To simplify notation, the four talkers selected for use in these intelligibility experiments are designated T1 through T4 (see Table 6.1). The

Table 6.1: Talker identification labels for the four talkers used in intelligibility tests.

Talker	Talker ID
SA	T1
RG	T2
MI	T3
EK	T4

speech of talkers T1 and T2 (SA and RG, respectively) was a superset of the speech analyzed in Chapters 3 and 4. The effect of each signal transformation on the acoustics of speech for these talkers was evaluated in Chapter 5. Similar measurements were made on the speech of T3 and T4 to verify that each signal transformation had the desired effect. The results of these measurements are reported in Appendix C and are similar to those reported for SA and RG.

Three speaking styles were presented at a normal speaking rate: conversational, clear, and processed conversational, and two speaking styles were presented at a slow speaking rate: conversational and processed conversational. The processed conversational style was achieved by applying the signal transformations described in Chapter 5 singly and in combination to conv/normal speech. The clear/normal condition was included to replicate Krause's[28] results for normal hearing listeners evaluating speech in noise and to verify that the intelligibility benefit extends to hearing impaired listeners in a quiet background. The slow rate processed condition was intended to consist of the signal transformations applied in combination, in order to determine if processed conversational speech at slow rates would approach the intelligibility of clear/slow speech.

To simplify notation, the signal transformations are designated as follows: Property A refers to the modification of formant frequencies (see Section 5.1), Property B refers to the modification of fundamental frequency (see Section 5.2), and Property C refers to the modification of temporal envelopes (see Section 5.3). In order to test these transformations singly and in combination, the database of conv/normal sentences allowed for 30 sentences to be tested per condition. Since 20 sentences re-

Table 6.2: The ten conditions tested for each talker. The order of presentation of test conditions was varied for each talker.

Speaking Mode	# of sentences
conversational/normal	30
clear/normal	30
processed(Property A)/normal	30
processed(Property B)/normal	30
processed(Property C)/normal	30
processed(Property A+B)/normal	20
processed(Property A+B+C)/normal	30
conversational/slow	30
processed(Property A+B)/slow	30
processed(Property A+B+C)/slow	30

mained, it was decided to test Property A+B as well, because informal listening tests indicated that this condition was the most intelligible of any possible pairing of signal transformations. This combination mode was also tested at the slow rate, with a full set of 30 sentences.

In total, ten conditions were examined, as shown in Table 6.2. The summation of one or more properties indicates that each signal transformation was applied in succession, beginning with the leftmost property.

6.2 Predicted Intelligibility Results

In Chapter 3, the speech-based STI measure was found to be highly correlated ($\rho = 0.9$) with the relative intelligibility of various speaking modes for T1 and T2. Therefore, the speech-based STI measure was applied to the stimuli in order to obtain a theoretical prediction of the relative intelligibility of each test condition. The procedure for calculating the speech-based STI was the same as that described in Chapter 3, except an SNR of -1.8dB was used in calculating the intensity envelope spectra for the noisy stimuli. The results are displayed in Figures 6-1 and 6-2.

For the data in Chapter 3, a range of 0.07 in STI corresponded to a range of 30 percentage points in intelligibility scores. Assuming this correlation extends to other talkers and speaking modes, the STI predicted an improvement in intelligibility over

conversational speech for most conditions at a normal speaking rate for talkers T1, T3, and T4. For T2, however, only clear/normal and processed(A)/normal styles showed a predicted improvement over conv/normal speech. Moreover, processed(C)/normal was predicted to have lower intelligibility than conv/normal speech for all talkers. This result was unexpected because this scheme was an enhancement of the temporal envelopes, believed to correspond to the measured increase in STI reported in Chapter 3. One possibility is that signal processing artifacts may have had a detrimental effect on intelligibility for this processing condition. At the slow speaking rate, the results were mixed, with both enhancement schemes predicting improved intelligibility over conv/slow speech for two talkers and degraded intelligibility for the other two talkers. Overall, these results suggest not only that significant intelligibility improvement should be expected for many of the test conditions but also that processing artifacts are not likely to prevent an observed intelligibility improvement for any conditions other than processed(C)/normal speech. This result is particularly important for the combination condition, since an artifact affecting high frequency modulations was identified in Chapter 5.

6.3 Experiments with Normal Hearing Subjects

Normal hearing listeners were employed to evaluate the intelligibility of the speech in the presence of additive wide-band noise. Intelligibility scores were based on the percentage of key words correct, using the scoring rules developed by Picheny *et al.*[45].

6.3.1 Listeners

Five normal hearing listeners (one male, four females) were obtained from the MIT community. The listeners were all native speakers of English who possessed at least a high school education. They ranged in age from 19 to 43 years. The results of each listener's hearing test is listed in Appendix D. Listeners were tested monaurally over TDH-39 headphones in a sound-treated room. Each listener selected the ear that

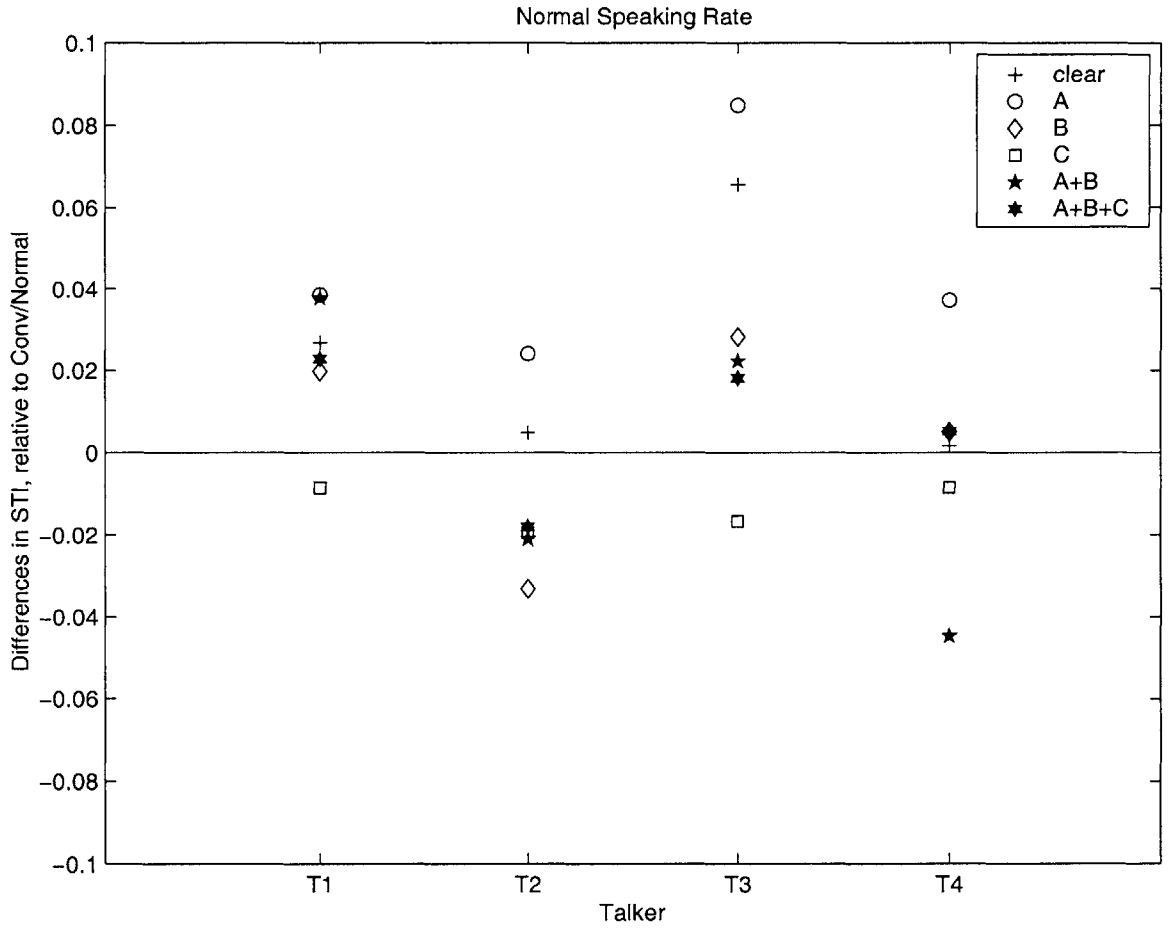

Figure 6-1: STI for all conditions at normal speaking rates relative to the STI for the conv/normal speech of each talker.

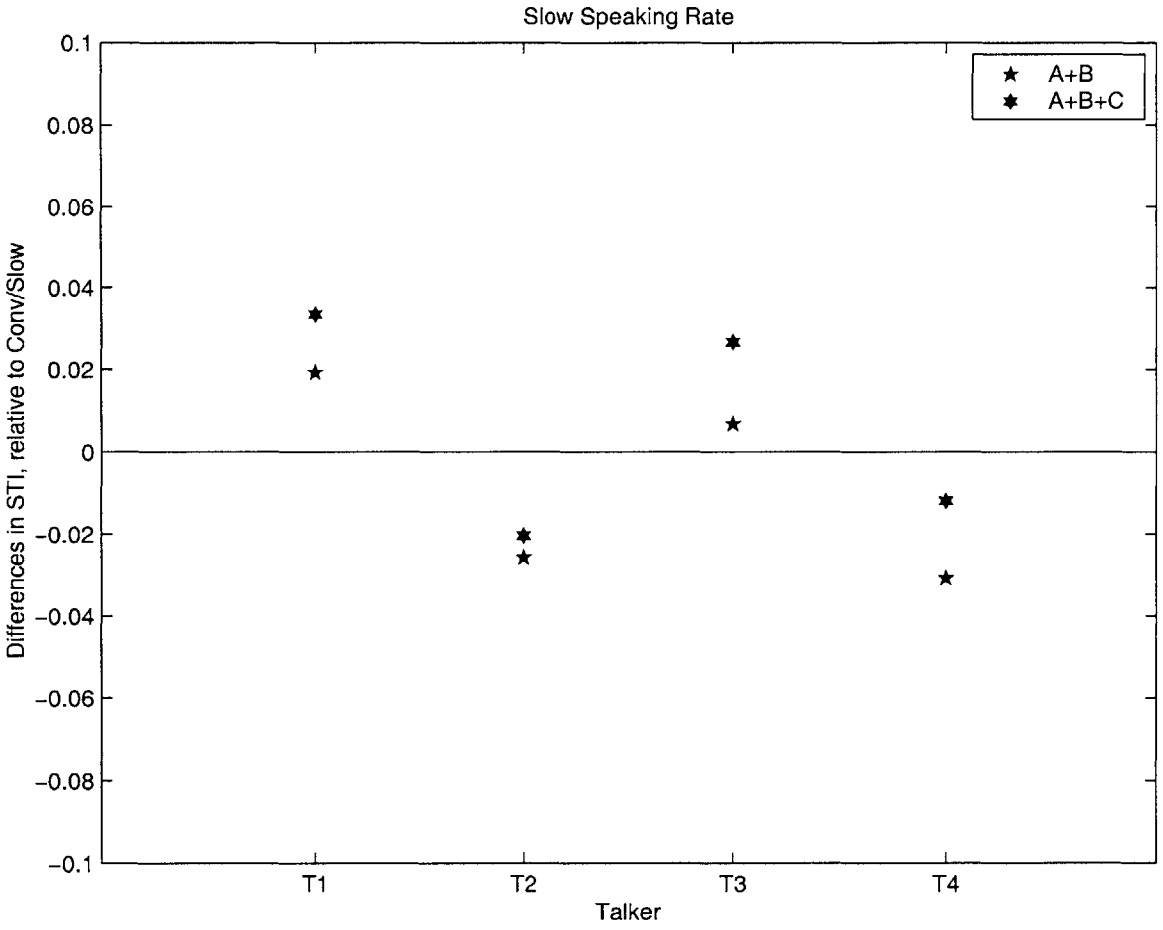

Figure 6-2: STI for all conditions at slow speaking rates relative to the STI for the conv/slow speech of each talker.

would receive the stimuli and was encouraged to switch the stimulus to the other ear when fatigued.

6.3.2 Presentation Sessions

Listeners were tested in five 150-minute sessions over the course of approximately two weeks. Listeners responded by writing their answers on paper. They were given as much time as needed to respond but were presented each sentence only once.

Listeners were presented a total of forty 30-sentence lists (4 talkers x 10 conditions/talker). In each session, listeners were tested on eight different lists. Every session included a brief break after the presentation of each list and a five-minute break after every other list. In addition, a 15-minute break was given near the halfway point of each session. Listeners were also encouraged to rest briefly as necessary.

6.3.3 Presentation Setup

The stimuli were stereo signals with speech on one channel and speech-shaped noise of the same rms level on the other channel. The speech-shaped noise samples were originally developed for the Hearing in Noise Test described by Nilsson *et al.*[39]. The waveforms were played from a PC through a DAL card. The PC was controlled by one of the listeners, who was seated at a terminal of the VAX. The speech was attenuated by 1.8 dB and added to the speech-shaped noise, resulting in a signal with SNR = −1.8 dB. All listeners were seated together in a sound-treated room and heard the same lists in the same order.

6.3.4 Results of Intelligibility Tests

For each of the talkers, the intelligibility scores (I), averaged across listeners, and the corresponding standard deviations (σ) are presented in Table 6.3. A t-test was applied, after an arcsine transformation ($\arcsin\sqrt{I_j/100}$) to equalize the variances, in order to determine the significance of difference between the mean of each test condition compared with conversational speech at the same speaking rate. Condi-

tions which were significant at the 0.01 level, after averaging scores across talker, are indicated by an asterisk. Intelligibility results for the individual listeners and corresponding t-tests are presented in Appendix E.

Table 6.3: Percent correct key-word scores (I) and corresponding standard deviations (σ) for each of the four talkers. Key-word scores are averaged across all five normal hearing listeners. Asterisks indicate I was significantly improved (p=0.01) over conversational I of the same speaking rate.

Mode		Talker				
		T1	T2	T3	T4	AVG
CONV/NORM	I	38.8	36.0	24.2	40.4	34.8
	σ	6.8	10.4	7.3	5.1	7.4
CLEAR/NORM	I	***66.3**	***47.0**	***46.2**	45.9	***51.3**
	σ	11.2	7.8	6.9	7.7	8.4
PROC(A)/NORM	I	43.1	42.0	***41.5**	37.3	***41.0**
	σ	10.8	7.6	8.8	5.8	8.2
PROC(B)/NORM	I	***54.5**	23.5	18.3	31.4	31.9
	σ	6.0	5.3	5.6	13.6	7.6
PROC(C)/NORM	I	31.8	17.9	10.2	24.0	21.0
	σ	10.0	5.9	1.7	8.5	6.5
PROC(A+B)/NORM	I	***55.4**	17.4	20.9	9.7	25.9
	σ	4.7	5.4	5.7	5.4	5.3
PROC(A+B+C)/NORM	I	29.5	8.6	6.5	8.7	13.3
	σ	5.9	5.5	3.6	4.3	4.8
CONV/SLOW	I	53.7	51.0	64.0	58.9	56.9
	σ	5.1	6.3	8.0	5.6	6.3
PROC(A+B)/SLOW	I	57.3	40.6	52.2	45.3	48.9
	σ	10.0	7.2	5.8	5.6	7.2
PROC(A+B+C)/SLOW	I	35.3	12.1	28.3	24.9	25.2
	σ	14.7	4.6	9.8	7.6	9.2

At normal speaking rates, clear speech was most intelligible (51%), followed in order of decreasing intelligibility by processed(A)/normal (41%), conv/normal (35%), processed(B)/normal (32%), processed(A+B)/normal (26%), processed(C)/normal (21%), and processed(A+B+C)/normal (13%). T-tests showed that the intelligibility advantage of clear/normal and processed(A)/normal speech relative to conv/normal speech was significant at the 0.01 level. The 16 percentage point improvement

for clear/normal over conv/normal is consistent with Krause[28], who measured a difference of 14 percentage points. At slow speaking rates, conversational speech was the most intelligible at 57% followed by processed(A+B)/slow (49%) and processed(A+B+C)/slow (25%). Conv/slow speech was the most intelligible overall, with an intelligibility advantage over conv/normal speech of 22 percentage points. This advantage was substantially larger than the 11 percentage point improvement observed in Krause's study[28].

The effect of talker and listener is shown in Figures 6-3 and 6-4, respectively. Some differences between talkers can be observed, such as the improvement observed for processed(B)/normal and processed(A+B)/normal relative to conv/normal for T1. T-tests verified that the scores in these two conditions relative to conv/normal speech were significant at the 0.01 level. No substantial differences were observed across listeners. The effect of interactions between talker and listener is shown in Figures 6-5 through 6-8.

An analysis of variance was performed on the normal rate conditions, after an arcsine transformation ($\arcsin\sqrt{I_j/100}$) to equalize the variances, in order to determine significant effects and interactions. Table 6.4 shows the results of this analysis with the factors talker, listener, and speaking mode. All three main factors as well as the talker x mode interaction were significant at the 0.05 level. The values of the F-distribution used for the F-tests were obtained from Bennet and Franklin[2]. Because the listener x mode interaction was not significant and did not account for any of the variance, this analysis confirms that the results are independent of listener.

The effect of mode was very strong, accounting for a large portion of the variance (45%). This effect may have resulted from a number of the conditions that were quite detrimental to intelligibility. Moreover, the talker x mode interaction did account for a substantial portion of the variance (12%). Therefore, to get more insight into whether the beneficial conditions were independent of talker as well as listener, two additional analyses were performed. In the first, an analysis of variance was performed on the data for only the clear/normal and conv/normal modes. The results of this analysis are shown in Figure 6.5. In this analysis, the same terms were significant

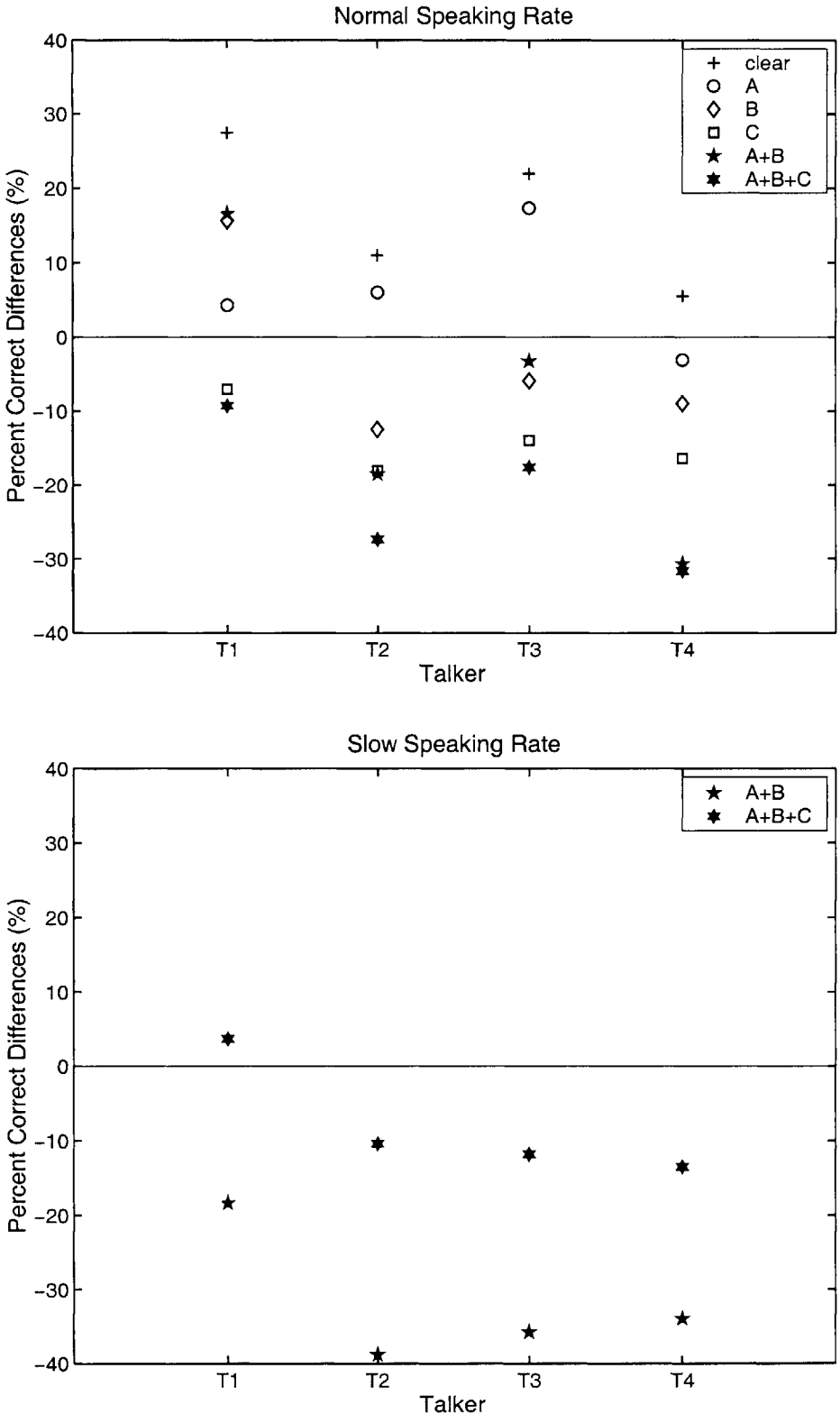

Figure 6-3: Percent correct scores relative to conversational mode for each talker, averaged across listener.

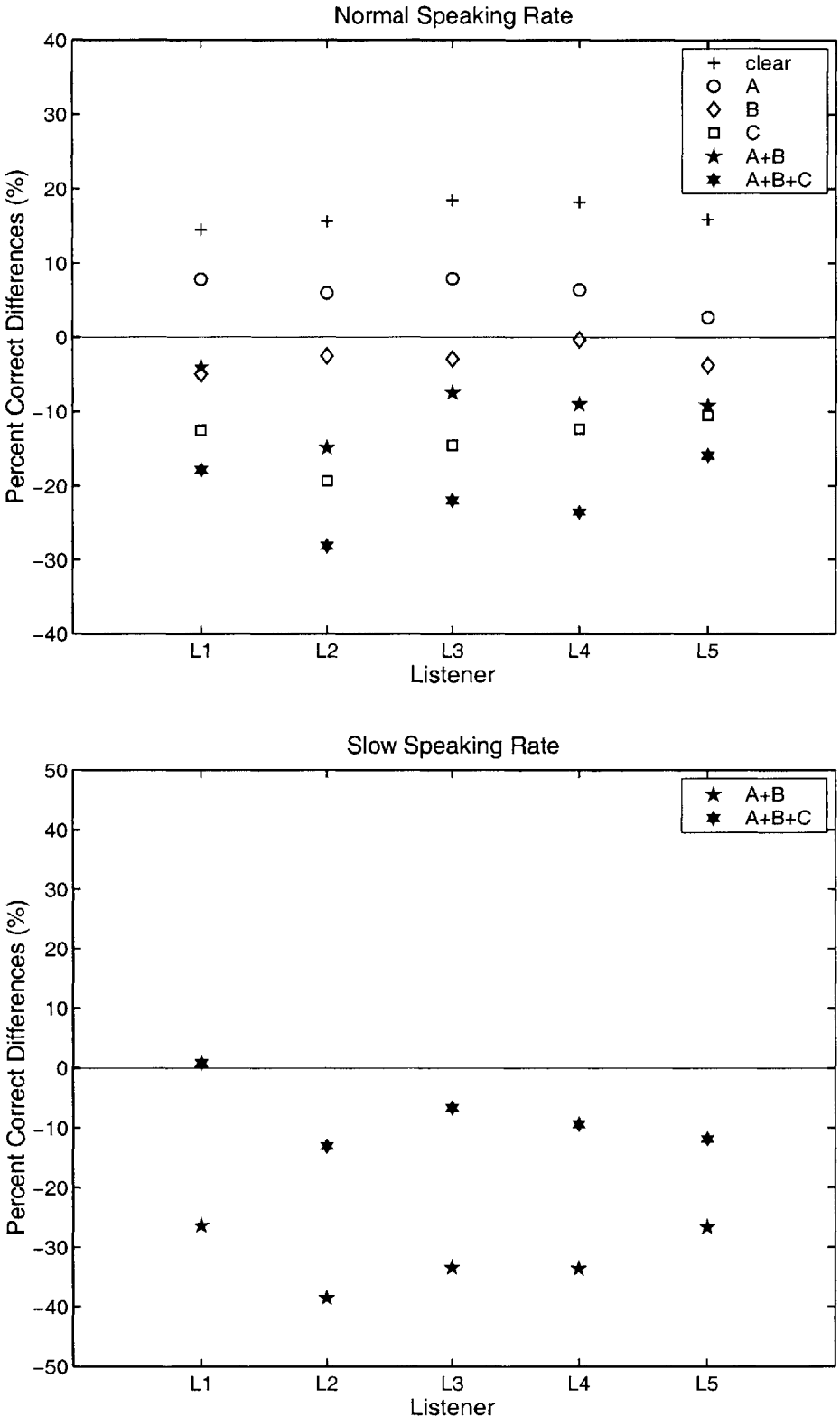

Figure 6-4: Percent correct scores relative to conversational mode for each listener, averaged across talker.

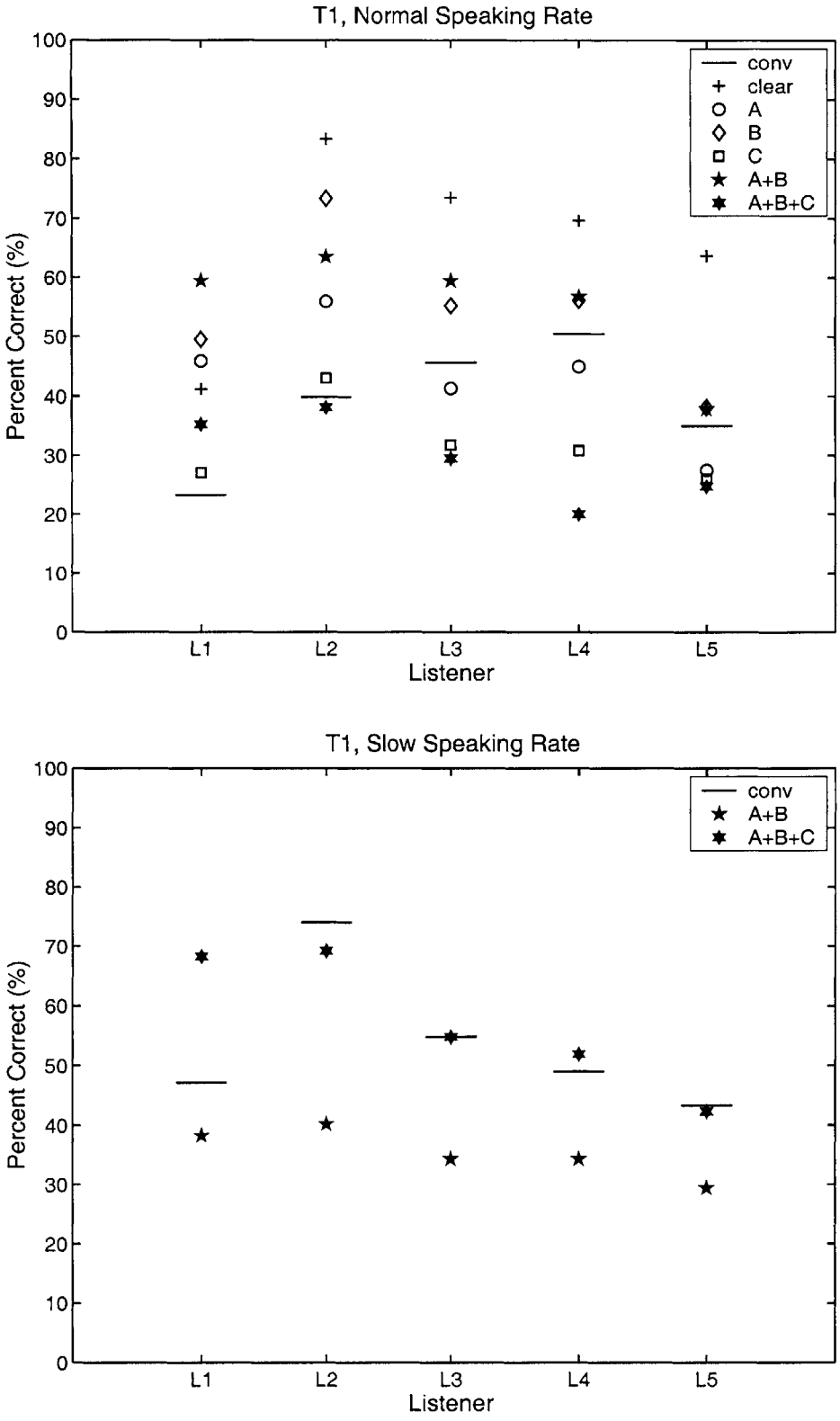

Figure 6-5: Percent correct scores, by listener, for T1 at normal and slow speaking rates.

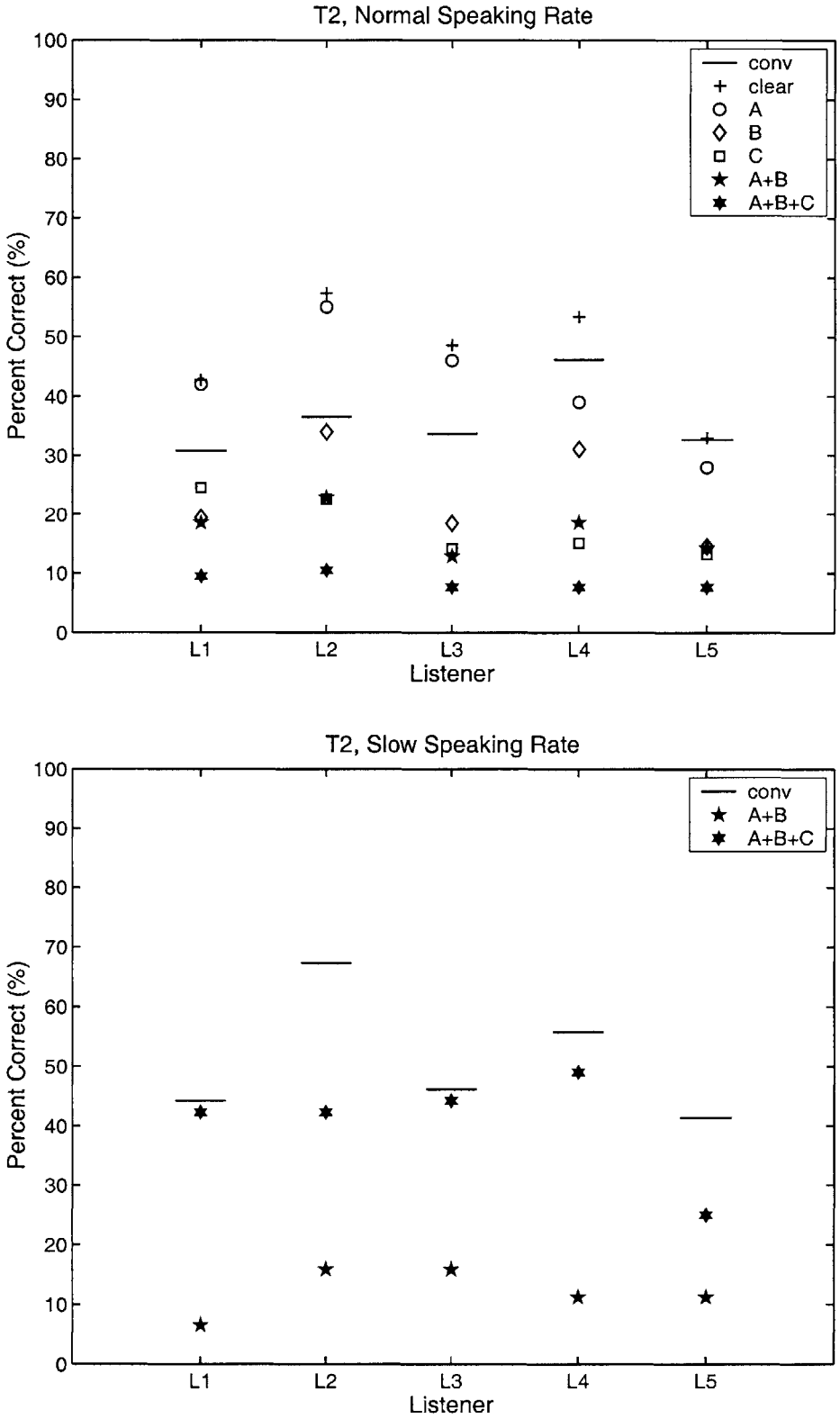

Figure 6-6: Percent correct scores, by listener, for T2 at normal and slow speaking rates.

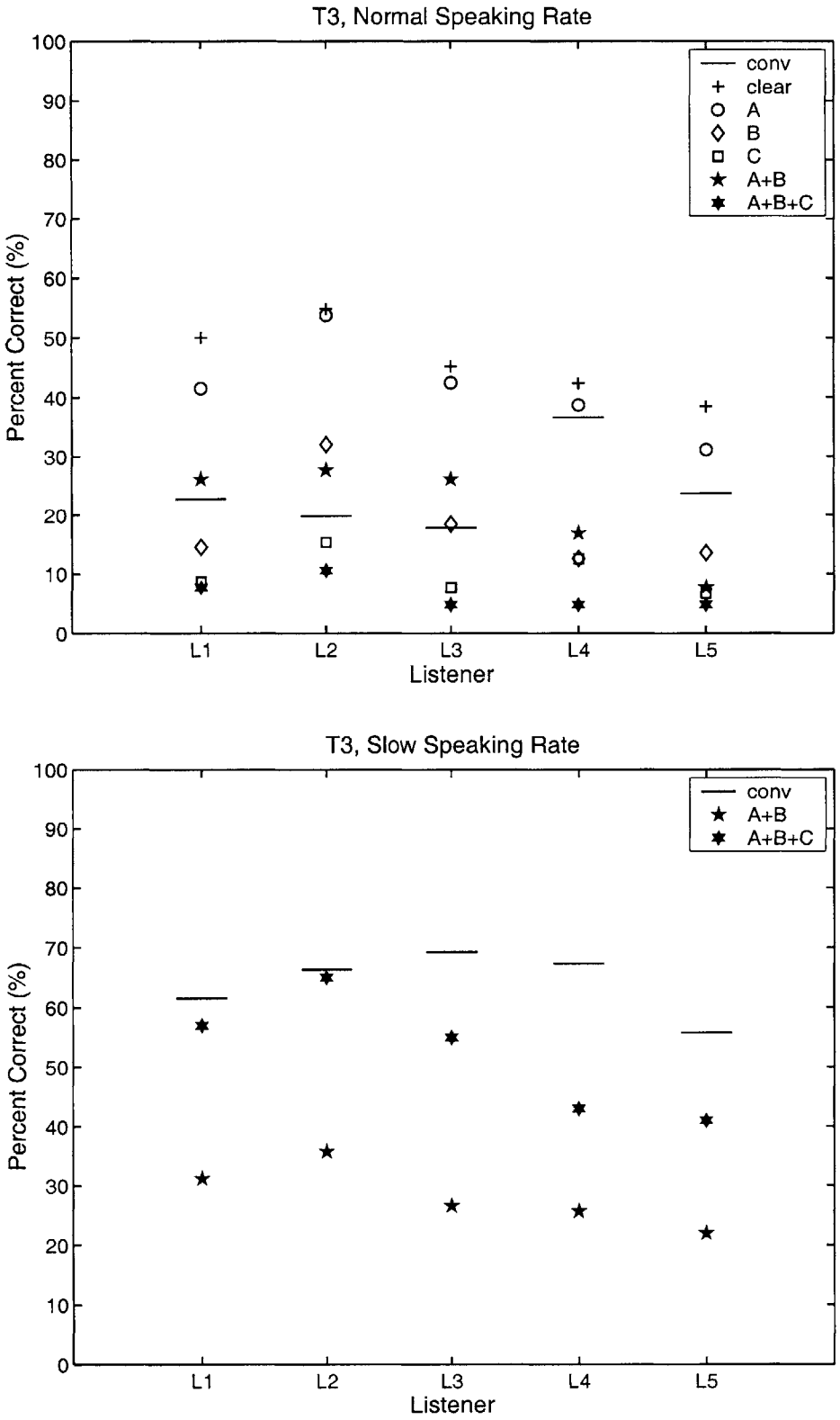

Figure 6-7: Percent correct scores, by listener, for T3 at normal and slow speaking rates.

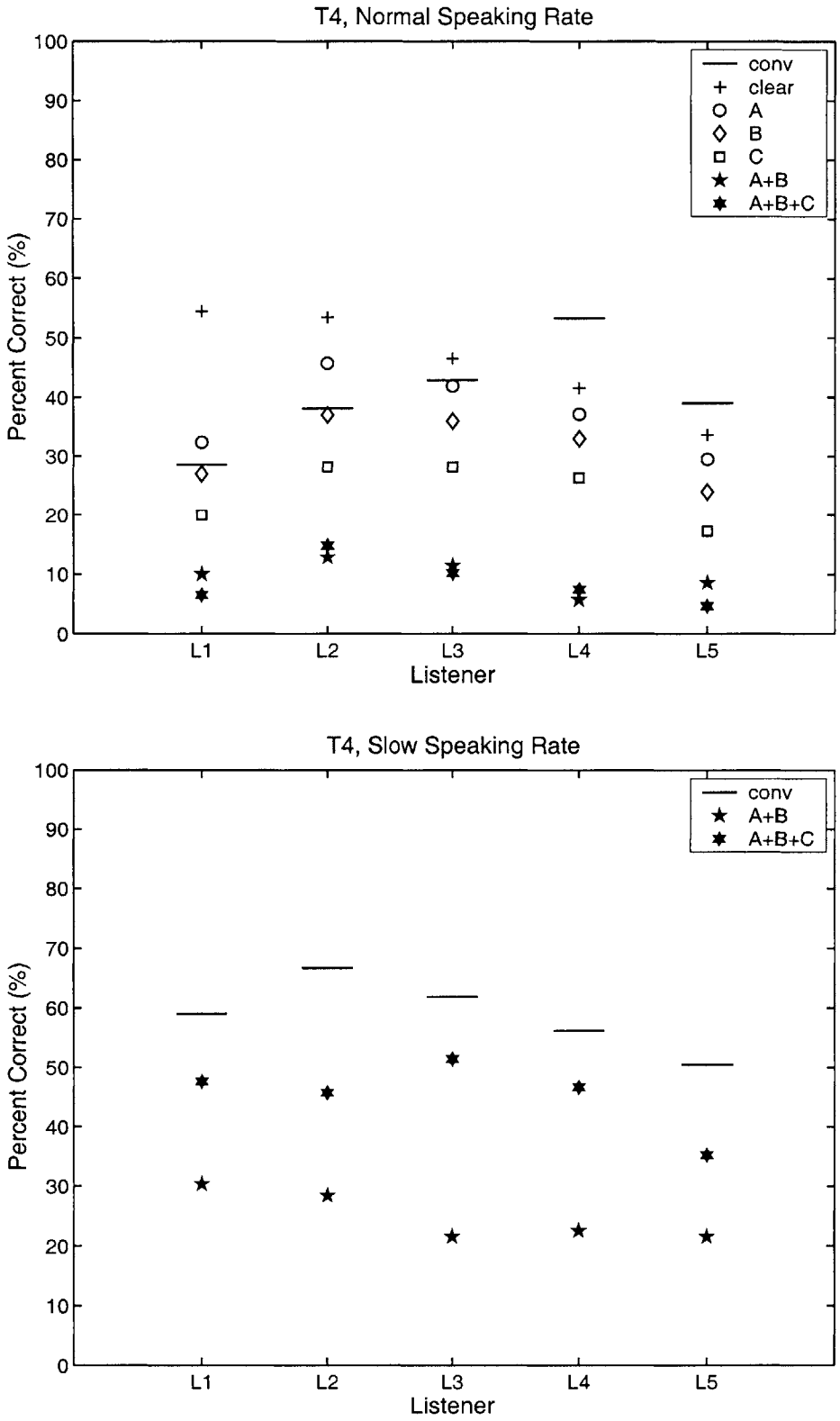

Figure 6-8: Percent correct scores, by listener, for T4 at normal and slow speaking rates.

Table 6.4: Analysis of variance of the intelligibility scores for the normal rate conditions. Factors which are significant at a 0.05 level are indicated by asterisks.

Sum of Squares	Mean Square	Degrees of Freedom	F-ratio	$\%\omega^2$		FACTOR
0.1804190	0.090209	2	-	0.8		REPS (R)
1.6096367	0.402409	4	44.9	7.9	*	LISTENER (L)
0.0506062	0.006326	8	-	0.0		LxR
4.0162821	1.338761	3	95.1	20.1	*	TALKER (T)
0.0995797	0.016597	6	-	0.2		TxR
0.1689998	0.014083	12	1.6	0.3		TxL
0.1622915	0.006762	24	-	0.0		TxLxR
8.3400154	1.390003	6	98.7	41.7	*	MODE (M)
0.1759119	0.014659	12	-	0.3		MxR
0.2336403	0.009735	24	1.1	0.0		MxL
0.2180989	0.004544	48	-	0.0		MxLxR
2.3049946	0.128055	18	13.2	10.8	*	MxT
0.8343142	0.023175	36	-	2.6		MxTxR
0.6988087	0.009706	72	1.1	0.2		MxTxL
0.7861647	0.005459	144	-	0.0		MxTxLxR
19.8797646	0.047446	419				TOTAL
2.5073861	0.008955	280				Residual (Error Term)

at the 0.05 level, with the addition of the talker x listener interaction. Since the talker x mode term remained significant and the listener x mode interaction remained insignificant, it can be concluded that the intelligibility advantage of clear/normal speech is independent of listener but somewhat dependent on talker.

Additional analyses of variance were performed for the two enhancement conditions which showed a significant intelligibility advantage over conv/normal speech in certain situations. Figure 6.6 shows the results for including only the conv/normal and processed(A)/normal speech, while Figure 6.7 shows the results for including only the conv/normal and processed(B)/normal speech. In both cases, the main factors as well as the mode x talker interaction were significant at the 0.05 level. Since the talker x mode interaction is significant in both these analyses, it can be concluded that the benefit derived from these enhancements is not entirely independent of talker. This is

Table 6.5: Analysis of variance of the intelligibility scores for conv/normal and clear/normal conditions. Factors which are significant at a 0.05 level are indicated by asterisks.

Sum of Squares	Mean Square	Degrees of Freedom	F-ratio	$\%\omega^2$		FACTOR
0.0326177	0.016309	2	-	0.0		REPS (R
0.5849844	0.146246	4	14.0	14.7	*	LISTENER (L)
0.0442439	0.005530	8	-	0.0		LxR
0.5866502	0.195550	3	9.7	15.0	*	TALKER (T)
0.1419902	0.023665	6	-	2.1		TxR
0.2425295	0.020211	12	1.9	3.1	*	TxL
0.1464131	0.006101	24	-	0.0		TxLxR
0.9596270	0.959627	1	314	25.7	*	MODE (M)
0.0687864	0.034393	2	-	1.3		MxR
0.0122205	0.003055	4	0.3	0.0		MxL
0.0438287	0.005479	8	-	0.0		MxLxR
0.2935438	0.097848	3	6.9	7.0	*	MxT
0.1917689	0.031961	6	-	3.5		MxTxR
0.1694965	0.014125	12	1.3	1.2		MxTxL
0.1685862	0.007024	24	-	0.0		MxTxLxR
3.6872873	0.030986	119				TOTAL
0.8382351	0.010478	80				Residual (Error Term)

Table 6.6: Analysis of variance of the intelligibility scores for conv/normal and processed(A)/normal conditions. Factors which are significant at a 0.05 level are indicated by asterisks.

Sum of Squares	Mean Square	Degrees of Freedom	F-ratio	$\%\omega^2$		FACTOR
0.0253761	0.012688	2	-	0.4		REPS (R
0.6559318	0.163983	4	19.0	31.2	*	LISTENER (L)
0.0422652	0.005283	8	-	0.0		LxR
0.1404355	0.046812	3	8.4	5.8	*	TALKER (T)
0.0706230	0.011771	6	-	0.9		TxR
0.0671522	0.005596	12	0.6	0.0		TxL
0.1735806	0.007233	24	-	0.0		TxLxR
0.1510084	0.151008	1	39.3	7.2	*	MODE (M)
0.0573619	0.028681	2	-	2.0		MxR
0.0153826	0.003846	4	0.4	0.0		MxL
0.0288325	0.003604	8	-	0.0		MxLxR
0.1918755	0.063958	3	10.9	8.3	*	MxT
0.1630304	0.027172	6	-	5.6		MxTxR
0.0704151	0.005868	12	0.7	0.0		MxTxL
0.1276591	0.005319	24	-	0.0		MxTxLxR
1.9809299	0.016646	119				TOTAL
0.6887288	0.008609	80				Residual (Error Term)

particularly true for processed(B)/normal speech, since the mode x talker interaction accounts for a higher proportion (12%) of the total variance than the mode factor (1%) alone.

6.4 Experiments with Hearing Impaired Subjects

Hearing-impaired listeners were tested to evaluate the intelligibility of the speech in a quiet background. As in the experiments with normal hearing listeners, intelligibility scores were based on the percentage of key words correct, using the scoring rules developed by Picheny *et al.*[45].

Table 6.7: Analysis of variance of the intelligibility scores for conv/normal and processed(B)/conditions conditions. Factors which are significant at a 0.05 level are indicated by asterisks.

Sum of Squares	Mean Square	Degrees of Freedom	F-ratio	$\%\omega^2$		FACTOR
0.0410775	0.020539	2	-	0.7		REPS (R
0.6687184	0.167180	4	20.3	19.5	*	LISTENER (L)
0.0130239	0.001628	8	-	0.0		LxR
1.2397494	0.413250	3	34.2	37.2	*	TALKER (T)
0.1288554	0.021476	6	-	2.4		TxR
0.1449919	0.012083	12	1.4	1.3		TxL
0.1336967	0.005571	24	-	0.0		TxLxR
0.0446147	0.044615	1	15.5	1.1	*	MODE (M)
0.0407063	0.020353	2	-	0.7		MxR
0.0115054	0.002876	4	0.3	0.0		MxL
0.0551079	0.006888	8	-	0.0		MxLxR
0.4159105	0.138637	3	29.5	12.0	*	MxT
0.1562786	0.026046	6	-	3.2		MxTxR
0.0563368	0.004695	12	0.6	0.0		MxTxL
0.1056138	0.004401	24	-	0.0		MxTxLxR
3.2561872	0.027363	119				TOTAL
0.6743601	0.008430	80				Residual (Error Term)

Table 6.8: Audiometric characteristics of the hearing-impaired listeners

Listener	Sex	Age	Test ear	WDS (% correct)	Thresholds (dB HL) at Frequency (Hz)					
					250	500	1000	2000	4000	8000
GI	M	65	Left	100	55	60	45	45	55	85
RK	M	64	Left	92	10	20	40	60	65	NR
GT	M	40	Right	100	50	55	55	60	90	85

6.4.1 Listeners

Three hearing-impaired listeners (all male) with stable sensorineural hearing losses participated in the experiment. The listeners were all native speakers of English who possessed at least a high school education. They ranged in age from 40 to 65 years. Listeners were tested monoaurally over headphones in a sound-treated room. The audiometric characteristics of the test ear for each listener are summarized in Table 6.8.

6.4.2 Presentation Sessions

Each listener was tested individually in four to six sessions. Sessions were two to three hours in duration. Listeners responded by writing their answers on paper. They were given as much time as needed to respond but were presented each sentence only once.

Listeners were presented a total of forty 30-sentence lists (4 talkers x 10 conditions/talker). In each session, listeners were tested on seven to twelve different lists. Every session included a brief break after the presentation of each list, and listeners were also encouraged to rest for longer periods as necessary.

6.4.3 Presentation Setup

Listeners were tested monaurally in a sound-treated room. For each listener, a frequency-gain characteristic was obtained using the NAL procedure[5] and implemented using a third-octave filter bank (General Radio, 1925). The waveforms were played from a PC through a DAL card. The PC was controlled by the listener, who

was seated at a terminal of the VAX. The listener was provided with an attenuator that could be used to adjust the overall system gain and was given the opportunity to make adjustments to the level at the beginning of each condition. This adjustment procedure ensured that the presentation level of each condition was as comfortable and audible as possible for each listener.

6.4.4 Results of Intelligibility Tests

For each of the talkers, the intelligibility scores (I), averaged across listeners, and the corresponding standard deviations (σ) are presented in Table 6.9. After an arcsine transformation, a t-test was applied to test the significance of differences between the mean of each test condition compared with conversational speech at the same speaking rate. No conditions exhibited an improvement over conversational speech that was significant at the 0.01 level, when scores were averaged across listeners. For T3, processed(A)/normal speech exhibited an improvement over conversational speech that was significant at the 0.05 level, as indicated in Table 6.9. The intelligibility results for individual listeners and corresponding t-test results are listed in Appendix E.

At normal speaking rates, clear speech was most intelligible at 69%, followed in order of decreasing intelligibility by conv/normal (62%), processed(A)/normal (61%), processed(B)/normal (54%), processed(A+B)/normal (52%), processed(C)/normal (28%), and processed(A+B+C)/normal (23%). T-tests showed no condition, when averaged across listeners, to have a significant advantage over conv/normal speech. At slow speaking rates, conversational speech was the most intelligible at 71% followed by processed(A+B)/slow (66%) and processed(A+B+C)/slow (29%). Conv/slow speech was the most intelligible overall, with an intelligibility advantage over conv/normal speech of 9 percentage points.

With the exception of the processed(A)/normal style, which did not provide as large of an overall benefit to hearing impaired listeners, the relative intelligibility of conditions for hearing-impaired listeners roughly parallel those of normal hearing listeners in noise. However, substantial differences can be observed between both talkers and listeners. These effects are shown in Figures 6-9 and 6-10, respectively.

Table 6.9: Percent correct key-word scores (I) and corresponding standard deviations (σ) for each of the four talkers. Key-word scores are averaged across all three hearing-impaired listeners. Asterisks indicate I was significantly improved (p=0.05) over conversational I of the same speaking rate.

Mode		Talker				
		T1	T2	T3	T4	AVG
CONV/NORM	I	80.6	52.9	48.2	67.9	62.4
	σ	3.4	7.2	5.0	4.1	4.9
CLEAR/NORM	I	79.1	68.6	58.3	68.6	68.7
	σ	6.8	7.9	7.0	5.7	6.8
PROC(A)/NORM	I	61.5	61.0	***61.9**	61.3	61.4
	σ	8.4	9.6	6.4	6.4	7.7
PROC(B)/NORM	I	73.0	53.4	46.3	43.4	54.0
	σ	10.4	9.6	8.2	15.8	11.0
PROC(C)/NORM	I	42.1	23.3	20.5	24.8	27.7
	σ	13.4	6.0	7.3	6.5	8.3
PROC(A+B)/NORM	I	80.2	49.5	49.5	27.1	51.6
	σ	4.5	7.4	4.9	2.0	4.7
PROC(A+B+C)/NORM	I	40.3	15.9	17.0	17.6	22.7
	σ	10.9	6.7	5.2	5.1	7.0
CONV/SLOW	I	76.3	65.4	78.2	62.9	70.7
	σ	7.3	7.6	9.4	12.5	9.1
PROC(A+B)/SLOW	I	73.4	60.9	62.3	67.3	66.0
	σ	6.9	6.4	9.3	6.8	7.3
PROC(A+B+C)/SLOW	I	37.6	19.6	28.1	31.4	29.2
	σ	8.7	4.5	6.5	6.6	6.6

Interactions between talker and listener are shown in Figures 6-11 through 6-14.

An analysis of variance was performed on the normal rate conditions, after an arcsine transformation (arcsin $\sqrt{I_j/100}$) to equalize the variances, in determining the significance of effects and interactions. Table 6.10 shows the results of this analysis with the factors talker, listener, and speaking mode. All three main factors as well as the talker x mode interaction were significant at the 0.05 level. Because the neither the listener x mode interaction nor the talker x mode interaction accounted for any of the variance, this analysis suggests that the results are independent of listener and talker.

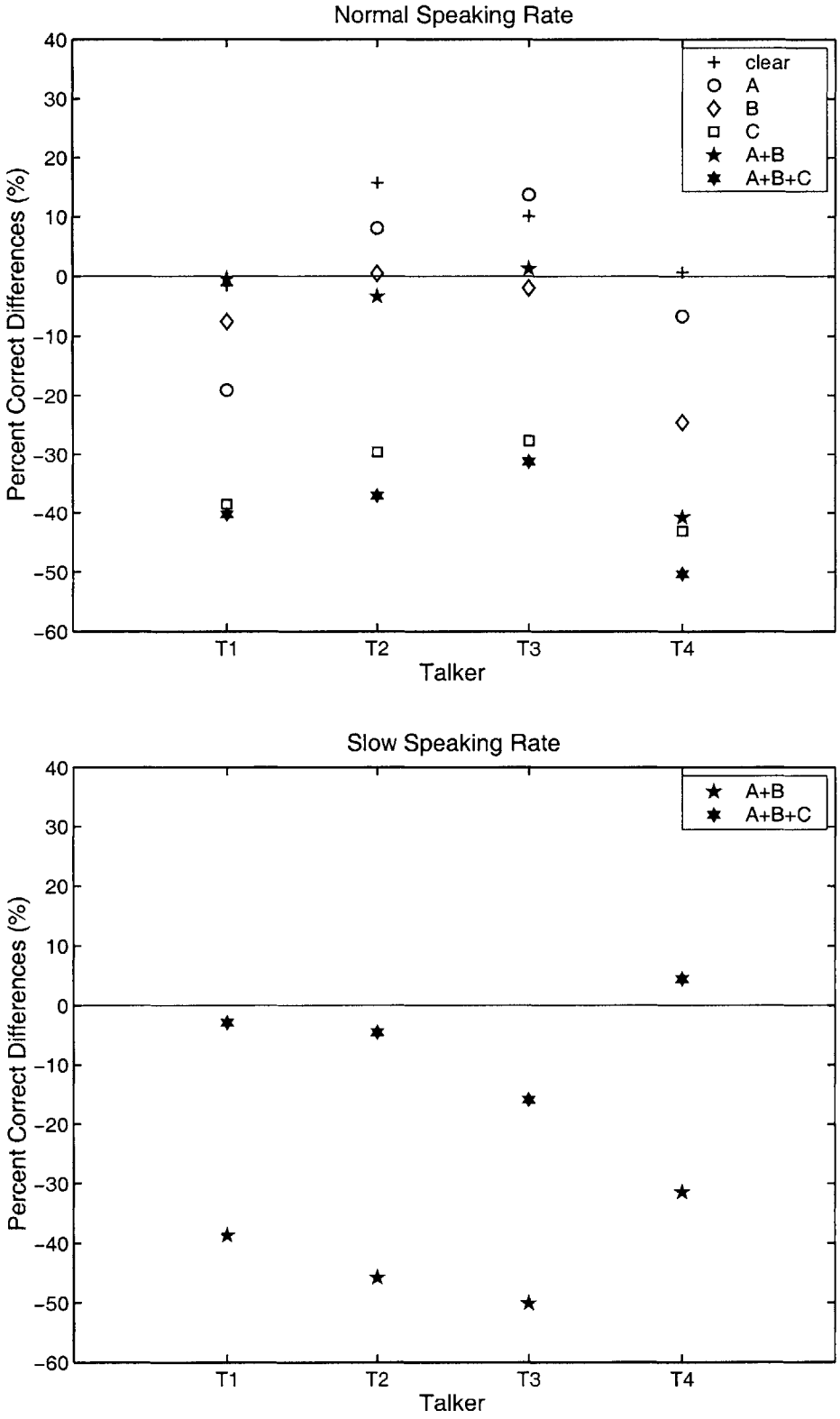

Figure 6-9: Percent correct scores relative to conversational mode for each talker, averaged across listener.

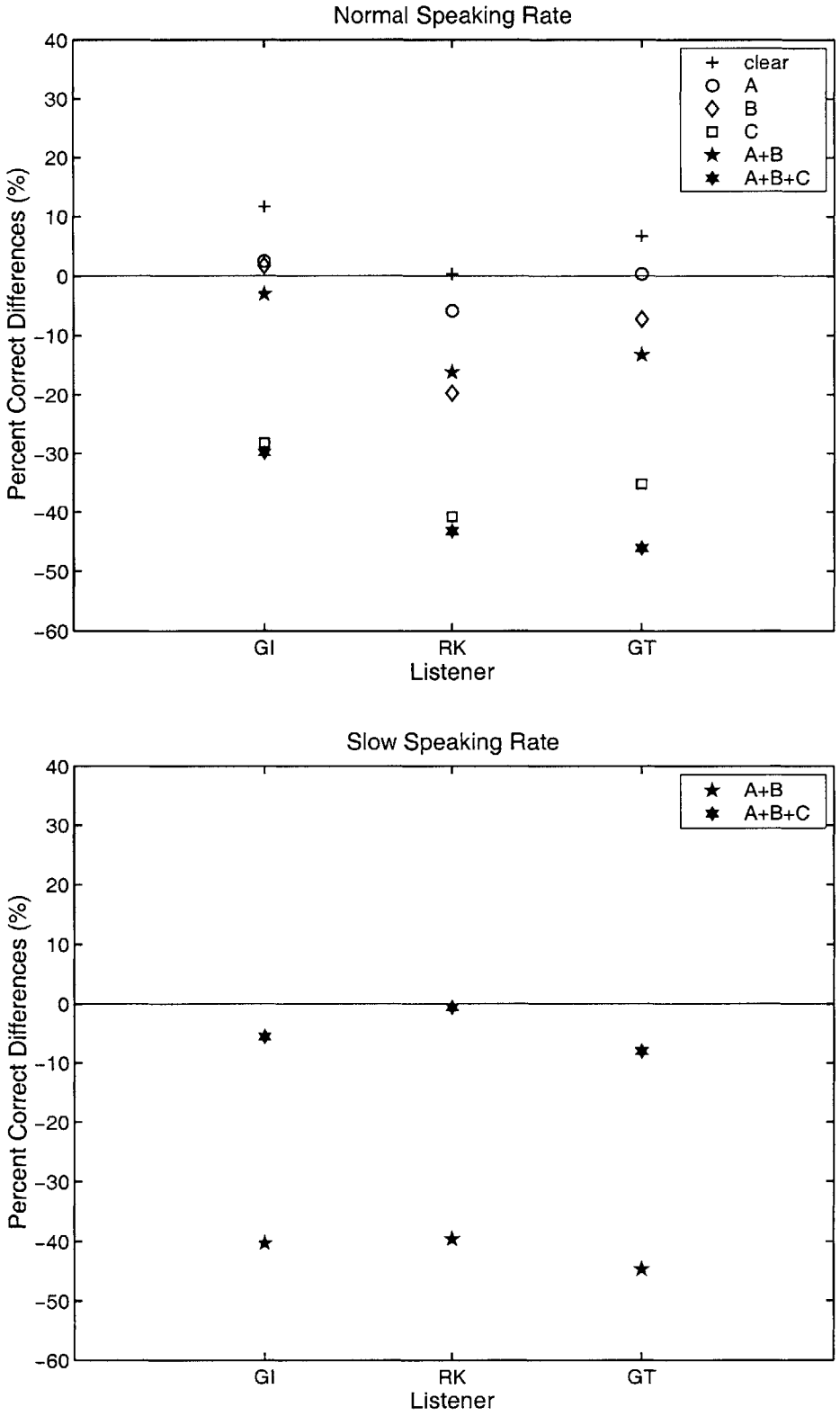

Figure 6-10: Percent correct scores relative to conversational mode for each listener, averaged across talker.

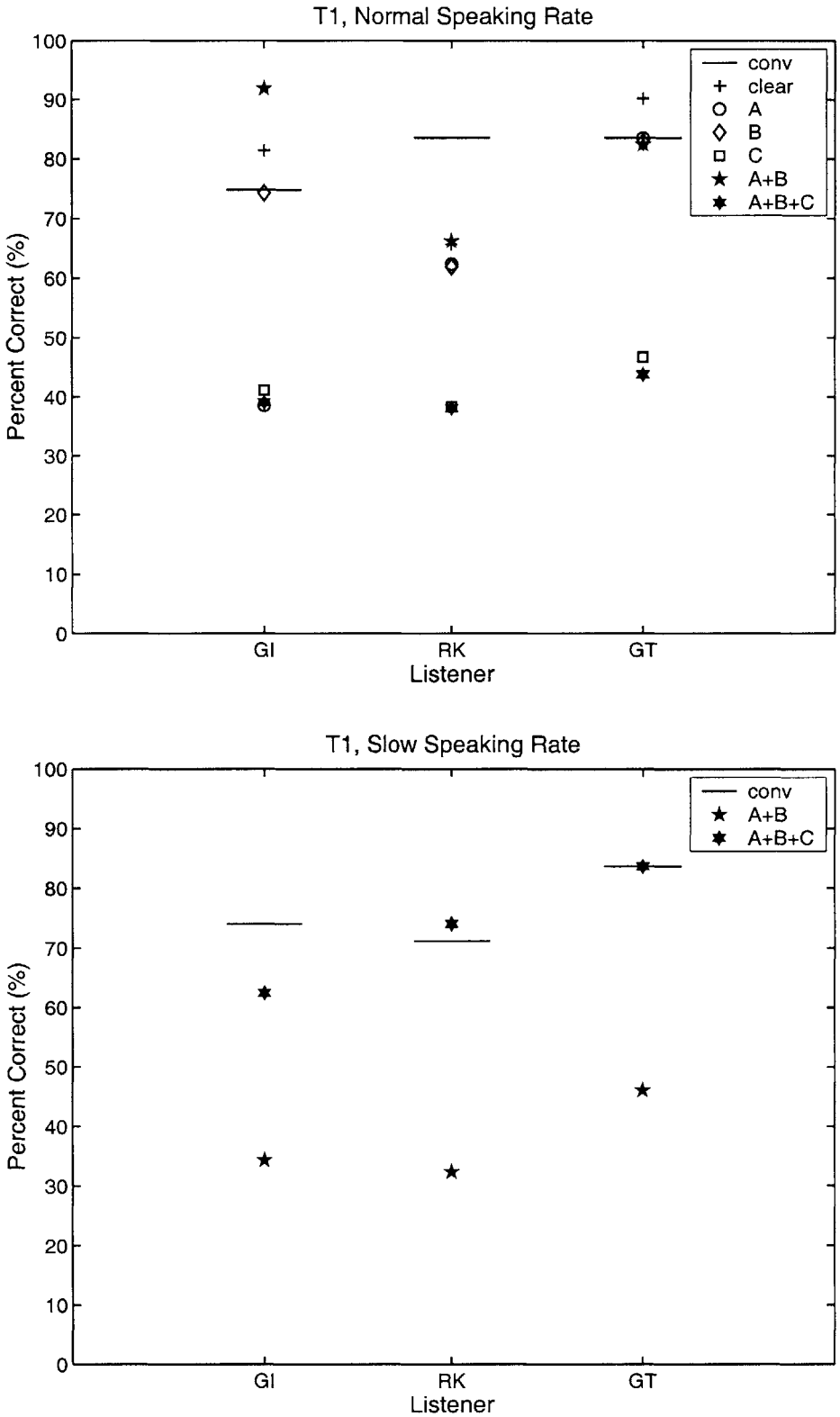

Figure 6-11: Percent correct scores, by listener, for T1 at normal and slow speaking rates.

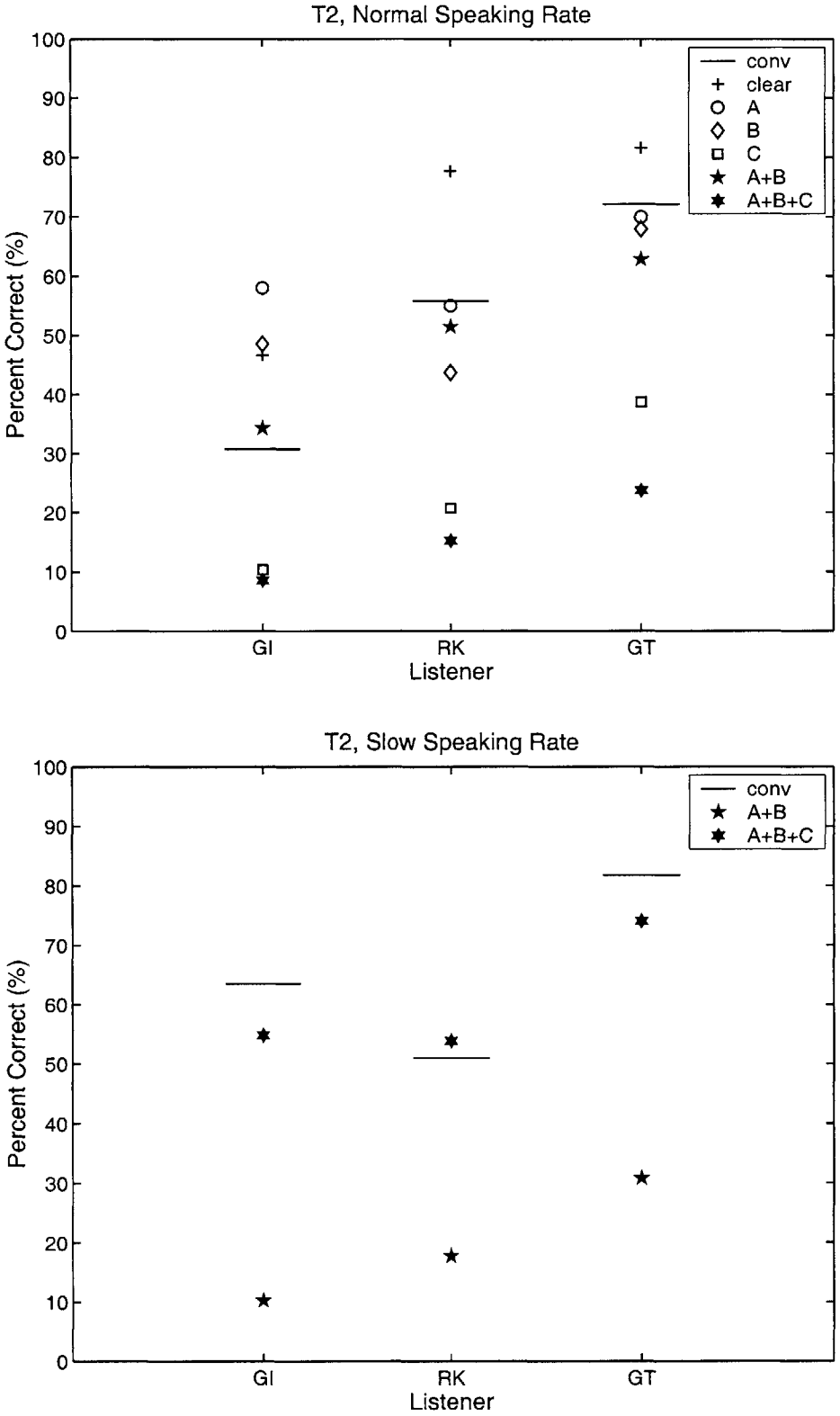

Figure 6-12: Percent correct scores, by listener, for T2 at normal and slow speaking rates.

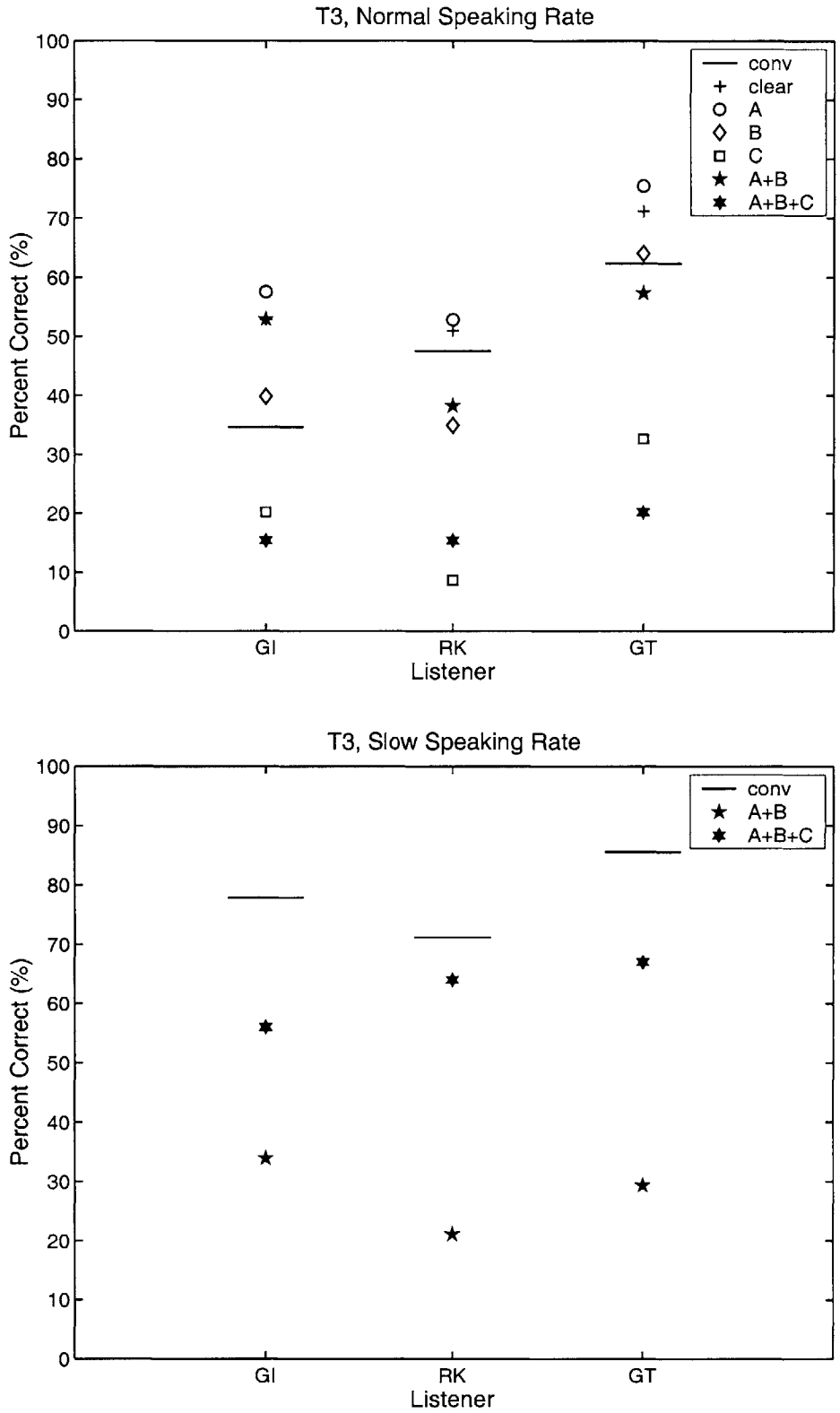

Figure 6-13: Percent correct scores, by listener, for T3 at normal and slow speaking rates.

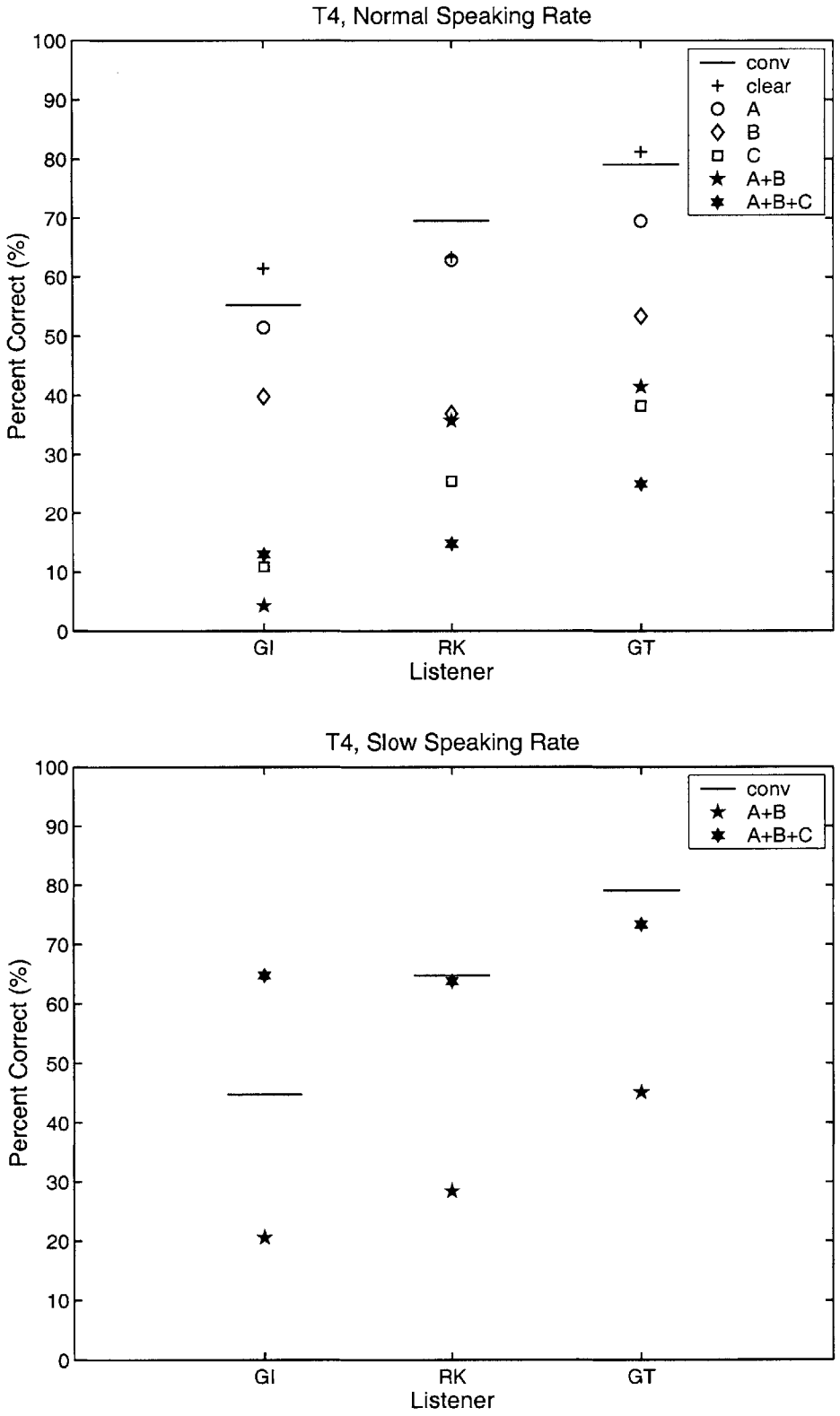

Figure 6-14: Percent correct scores, by listener, for T4 at normal and slow speaking rates.

Two additional analyses of variance were performed on conditions that seemed to produce some intelligibility advantage over conv/normal speech in certain situations. In one case, an analysis of variance was performed on the data for only the clear/normal and conv/normal modes. The results of this analysis are shown in Figure 6.11. In this analysis, the listener and talker main factors were significant at the 0.05 level, as well as the listener x talker, listener x mode, and listener x talker x mode interaction terms. The mode factor was not significant and only accounted for 3% of the variance, suggesting that reliable intelligibility benefits from clear/normal speech do not extend to hearing impaired listeners. However, Figures 6-11 through 6-14 and the individual scores listed in Appendix E do show that clear/normal speech provided a significant benefit to some listeners for some talkers. One explanation for this result is that the differing strategies talkers use to obtain clear/normal speech may vary in effectiveness when combined with the specific audiometric characteristics of each hearing impaired listener.

An additional analysis of variance was performed (see Table 6.12) including only the conv/normal and processed(A)/normal speech, since this enhancement appeared to be beneficial in certain situations. Again, the listener and talker factors were significant, and the mode factor was not. The listener x talker x mode and the talker x mode interactions were also significant, and the talker x mode interaction accounted for 15% of the variance, suggesting that the benefit provided by the enhancement was somewhat dependent on talker. The benefits of this enhancement, however, do appear to be independent of listener, since the listener x mode interaction is not significant.

6.5 Summary

The speech-based STI predicted that a majority of the processed conditions would improve intelligibility over conv/normal speech presented in wideband noise to normal hearing listeners. Actual experiments with normal hearing listeners revealed an advantage only for clear/normal speech and processed(A)/normal speech. Two other conditions, processed(B)/normal and processed(A+B) normal, improved the intelli-

Table 6.10: Analysis of variance of the intelligibility scores for the normal rate conditions. Factors which are significant at a 0.05 level are indicated by asterisks.

Sum of Squares	Mean Square	Degrees of Freedom	F-ratio	$\%\omega^2$		FACTOR
0.0529974	0.026499	2	-	0.0		REPS (R
1.9920218	0.996011	2	11.6	10.5	*	LISTENER (L)
0.0564324	0.014108	4	-	0.0		LxR
2.6415355	0.880512	3	14.3	13.8	*	TALKER (T)
0.0341633	0.005694	6	-	0.0		TxR
0.3689643	0.061494	6	0.7	0.0		TxL
0.0302291	0.002519	12	-	0.0		TxLxR
8.2540960	1.375683	6	67.6	44.6	*	MODE (M)
0.1168053	0.009734	12	-	0.0		MxR
0.2440625	0.020339	12	0.8	0.0		MxL
0.1294419	0.005393	24	-	0.0		MxLxR
1.3980727	0.077671	18	3.1	0.0	*	MxT
0.5130246	0.014251	36	-	0.0		MxTxR
0.9076924	0.025214	36	0.3	0.0		MxTxL
0.5086360	0.007064	72	-	0.0		MxTxLxR
17.2481766	0.068718	251				TOTAL
1.44173	0.085817	168				Residual (Error Term)

Table 6.11: Analysis of variance of the intelligibility scores for conv/normal and clear/normal conditions. Factors which are significant at a 0.05 level are indicated by asterisks.

Sum of Squares	Mean Square	Degrees of Freedom	F-ratio	$\%\omega^2$		FACTOR
0.0052471	0.002624	2	-	0.0		REPS (R)
0.7777805	0.388890	2	68.7	31.5	*	LISTENER (L)
0.0088930	0.002223	4	-	0.0		LxR
0.8544769	0.284826	3	8.6	34.4	*	TALKER (T)
0.0284688	0.004745	6	-	0.0		TxR
0.1983894	0.033065	6	5.8	6.8	*	TxL
0.0842514	0.007021	12	-	0.1		TxLxR
0.0868774	0.086877	1	3.4	3.3		MODE (M)
0.0149378	0.007469	2	-	0.1		MxR
0.0503829	0.025191	2	4.4	1.6	*	MxL
0.0091284	0.002282	4	-	0.0		MxLxR
0.0958253	0.031942	3	2.1	3.2		MxT
0.0153489	0.002558	6	-	0.0		MxTxR
0.0912916	0.015215	6	2.7	2.3	*	MxTxL
0.1056297	0.008802	12	-	1.5		MxTxLxR
2.4269290	0.034182	71				TOTAL
0.2719051	0.005665	48				Residual (Error Term)

Table 6.12: Analysis of variance of the intelligibility scores for conv/normal and processed(A)/normal conditions. Factors which are significant at a 0.05 level are indicated by asterisks.

Sum of Squares	Mean Square	Degrees of Freedom	F-ratio	$\%\omega^2$		FACTOR
0.0186802	0.009340	2	-	0.2		REPS (R)
0.8240768	0.412038	2	65.1	38.6	*	LISTENER (L)
0.0256983	0.006425	4	-	0.0		LxR
0.3604022	0.120134	3	17.8	16.3	*	TALKER (T)
0.0189501	0.003158	6	-	0.0		TxR
0.0404612	0.006744	6	1.1	0.1		TxL
0.0347014	0.002892	12	-	0.0		TxLxR
0.0023393	0.002339	1	0.2	0.0		MODE (M)
0.0737532	0.036877	2	-	2.9		MxR
0.0294930	0.014746	2	2.3	0.8		MxL
0.0165116	0.004128	4	-	0.0		MxLxR
0.3323220	0.110774	3	17.5	14.9	*	MxT
0.0209595	0.003493	6	-	0.0		MxTxR
0.2011036	0.033517	6	5.3	7.8	*	MxTxL
0.0943176	0.007860	12	-	0.9		MxTxLxR
2.0937700	0.029490	71				TOTAL
0.3035719	0.006324	48				Residual (Error Term)

gibility of only T1's conversational speech. Since these processing schemes involved manipulating F0, and T1 was the only male talker in the study, it is possible that this modification is only beneficial for male talkers. This result would not be entirely unexpected, since females have a wider F0 range than males, on average, and further widening of the range may sound unnatural. Another possibility is that the LPC analysis-synthesis used to modify F0 introduced artifacts that degraded intelligibility to a greater extent that the advantage provided by modifying F0. This idea will be explored in more detail in Chapter 7.

Hearing impaired listeners did not obtain an intelligibility benefit from clear/normal or processed(A)/normal speech as reliably as normal/hearing listeners in noise. These conditions, however, did provide a statistically significant benefit for some combinations of listeners and talkers. The fact that the benefit was not more robust across talkers and listeners most likely stems from the fact that the talkers had employed somewhat different strategies for producing clear/normal speech. Each hearing impaired listener had unique audiometric characteristics, and some of the styles of clear speech may have been better suited to those characteristics than others.

Chapter 7

Discussion

Despite the fact that the signal transformations applied in this thesis were based on the acoustics of clear/normal speech, intelligibility tests showed that the transformations provided no intelligibility improvement for hearing-impaired listeners and improvement in only a few conditions for normal hearing listeners in noise. While one explanation of these results was that processing artifacts resulted in reduced intelligibility, this should have been reflected in the STI measurements. Instead, the STI predicted that many of the signal transformations would improve intelligibility over conv/normal speech. Since this improvement was not observed, further acoustic measurements and an additional intelligibility experiment were completed in order to determine the effect of processing artifacts on both the acoustics and the intelligibility of the processed speech.

7.1 Additional Acoustic Analysis of Signal Transformations

During development of the signal transformation schemes, acoustic measurements verified that each transformation produced the desired acoustic manipulation of the speech signal (see Chapter 5). However, processing artifacts could have altered other properties of the speech, presumably resulting in reduced intelligibility. Because

Chapters 3 and 4 identified three acoustic properties associated with clear/normal speech and each transformation scheme was intended to modify only one of those properties, unintended alteration of either of the other two properties could be detrimental to intelligibility. Those properties were therefore examined for each transformation scheme, using the measurement procedures described in Chapter 3.

For processed(A)/normal speech, the transformation associated with modification of formant frequencies, no effect was found on either fundamental frequency distribution or temporal envelopes. Similarly, no change in fundamental frequency distribution or long-term spectra was measured for processed(C)/normal speech, the transformation associated with envelope modification. The results of these measurements are presented in Sections B.3 through B.5 of Appendix B and Sections C.4 through C.6 of Appendix C.

For processed(B)/normal speech, however, some change was measured in both the intensity envelope spectra and the long-term spectra of the processed speech relative to conversational speech, even though the LPC analysis-synthesis scheme was intended only to modify fundamental frequency. Figures 7-1 and 7-2 show the intensity envelope spectra for SA (T1) and RG (T2). While the envelope spectra of SA changed little as a result of processing, RG's processed spectra showed an increase in modulation depth for most modulation frequencies in the 125Hz, 250Hz, 4000Hz, and 8000Hz bands as well as a small increase in modulation depth for low frequency modulations in the 500Hz band. Similar changes in envelope spectra for T3 in the 250Hz band and for T4 in the 125Hz, 250Hz, and 1000Hz bands were observed and are shown in Figures C-12 and C-13 of Appendix C. Since the magnitude of the effect of processing on the envelope spectra varied substantially across talkers, the most likely explanation is that the LPC analysis-synthesis system varied in its ability to reproduce the specific vocal characteristics (creakiness, nasality, breathiness, etc.) of each talker. The envelope spectra data suggest that the LPC analysis-synthesis scheme was most successful in reproducing the vocal qualities for SA, moderately successful for T3, somewhat less successful for T4, and least successful for RG. This would not be unexpected, since some vocal characteristics are known to be difficult

for LPC processing to reproduce and can lead to reduced intelligibility after analysis-synthesis[26]. Thus, any vocal quality alterations introduced by LPC processing for T3, T4, and particularly RG, could have contributed to the reduced intelligibility of the processed(B)/normal condition for these talkers. This possibility is explored in the intelligibility experiment described in Section 7.2.

Figures 7-3 and 7-4 show the long-term spectra of processed(B)/normal speech relative to conv/normal speech for SA and RG, respectively. For both talkers, the LPC analysis-synthesis processing inadvertently resulted in 2–4dB boost in the mid- to high-frequency range. For SA, this boost somewhat larger and was present at all frequencies above 300Hz. For RG, frequencies were boosted between 300Hz and 3000Hz. Similar results were obtained for T3 and T4, shown in Figures C-8 and C-9 of Appendix C. Such an artifact is not likely to have degraded intelligibility. On the contrary, the analysis from Chapters 3 and 4 suggest this change in long-term spectra should be beneficial to intelligibility. However, in the intelligibility experiments of Chapter 6, processed(B)/normal speech was not more intelligible than conv/normal speech for any talker other than T1. Therefore, the LPC analysis-synthesis processing may have introduced additional artifacts that did have a detrimental effect on intelligibility. This possibility was addressed with an additional intelligibility experiment, described below.

7.2 Follow-up Intelligibility Experiment

In order to assess the effect of processing artifacts on intelligibility, a follow-up experiment was conducted. The experiment compared each processing scheme not only with conversational speech, as in the previous experiments, but also with speech that passed through the processing without modifying the specified acoustic parameter. This processed/unaltered speech should contain processing artifacts similar to the processed/enhanced speech but should be otherwise nearly identical acoustically to conversational speech.

While artifacts could have reduced intelligibility for any of the processing condi-

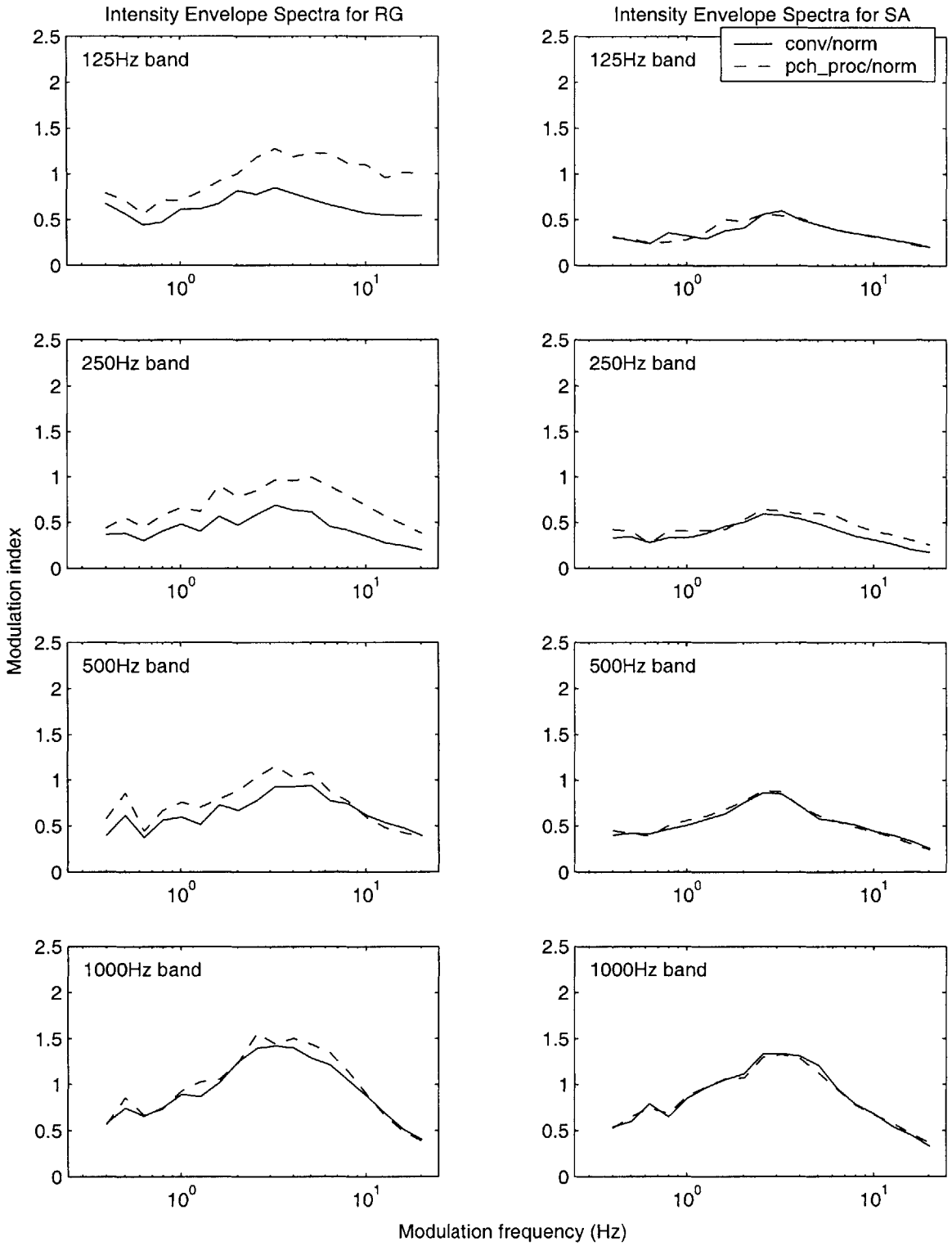

Figure 7-1: Spectra of intensity envelopes, before and after applying the pitch processing, for Talkers RG and SA in lower four octave bands.

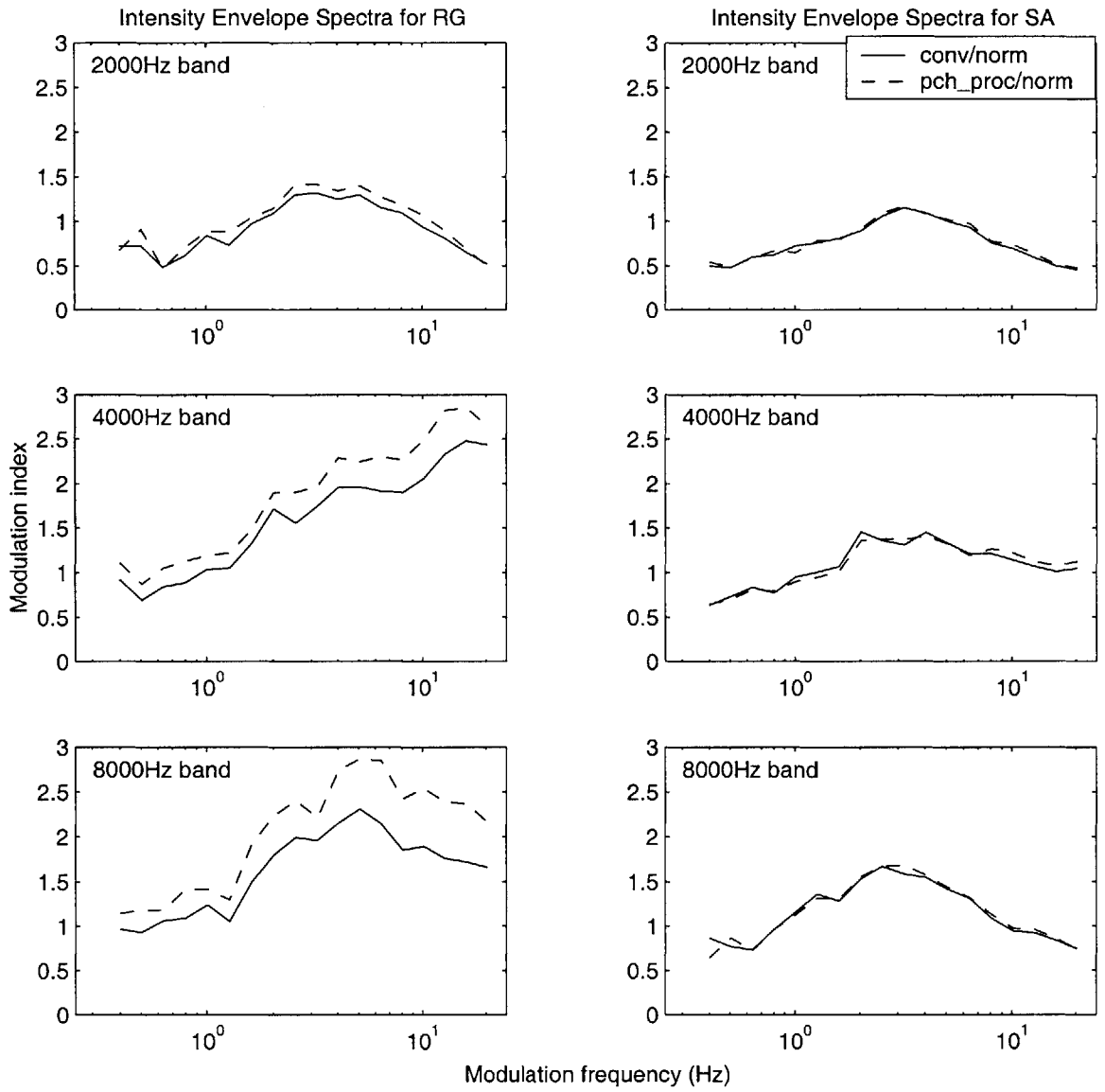

Figure 7-2: Spectra of intensity envelopes, before and after applying pitch processing, for Talkers RG and SA in upper three octave bands.

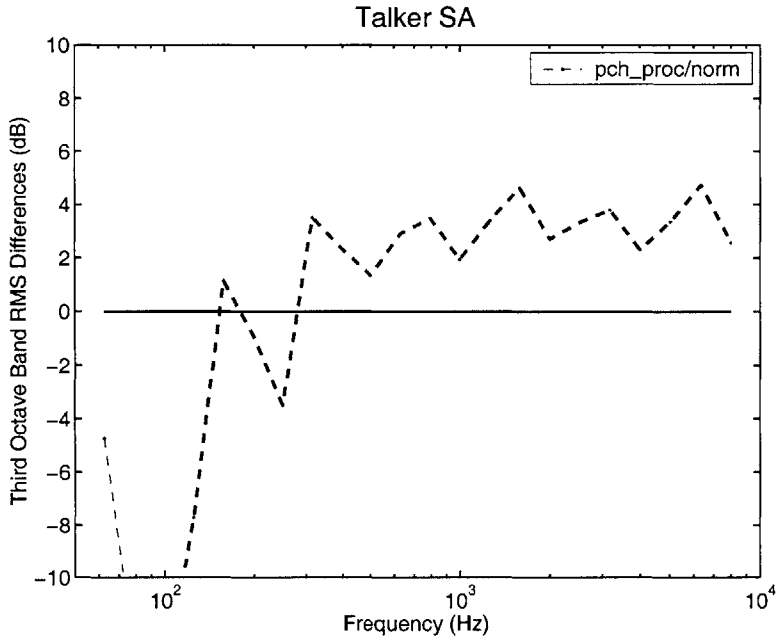

Figure 7-3: Third-octave band RMS spectral differences of SA's clear/normal and (pitch) processed/normal modes relative to conv/normal speech. A side effect of the pitch processing was high-frequency emphasis of the speech spectrum.

tions, the results of the intelligibility experiments in Chapter 6 suggested that processed(B)/normal and processed(C)/normal speech may have been most affected, since neither of these conditions provided an intelligibility benefit over conv/normal speech, when averaged across talker and listener. In addition, STI measurements also predicted that processed(C)/normal speech would be less intelligible than conv/normal speech, suggesting that the processing may have introduced artifacts. Additional acoustic measurements revealed that processed(B)/normal speech included a high-frequency boost that should have been associated with increased intelligibility. However, processed(B)/normal and processed(A+B)/normal speech were more intelligible than conversational speech only for T1, a male talker. An intelligibility advantage for these conditions was not observed for the other three talkers, who were female. Since these conditions involved a manipulation of F0, and T1 was the only male talker in the study, it is possible that this modification is only beneficial for male talkers. However, it is also possible that the LPC analysis-synthesis used to modify F0 may

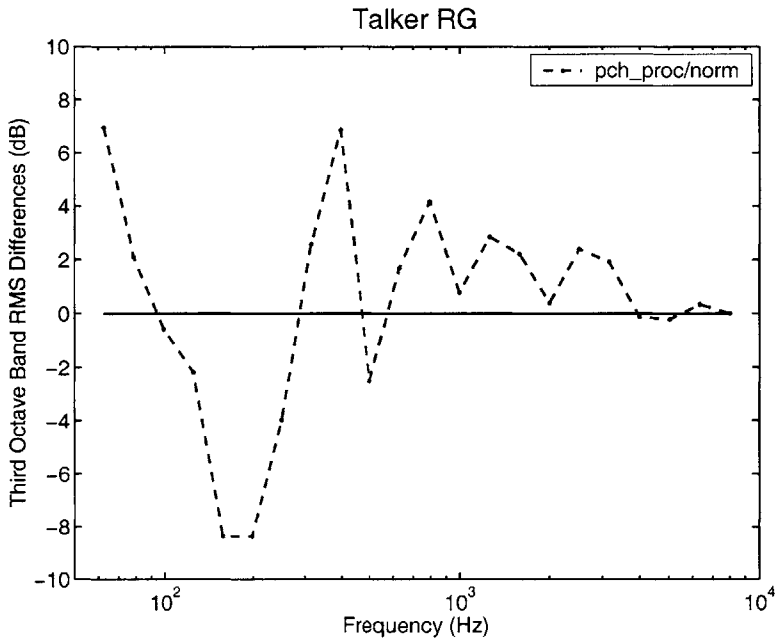

Figure 7-4: Third-octave band RMS spectral differences of RG's clear/normal and (pitch) processed/normal modes relative to conv/normal speech. A side effect of the pitch processing was high-frequency emphasis of the speech spectrum.

have introduced artifacts for female talkers that degraded intelligibility to a greater extent that the advantage provided by modifying F0. The follow-up intelligibility experiment was conducted to investigate this question as well as the effects of artifacts on the other processing schemes.

7.2.1 Methods

For each talker, seven conditions (see Table 7.1) were tested: one conversational and six processed conditions. Although 30 sentences were presented in six of these conditions, only 20 sentences could be tested in the remaining condition due to the total number of sentences in the database. The processed(enhanced_formant) condition was selected, since this was already shown to be significantly more intelligible than conversational speech for normal hearing listeners in noise in the previous experiment. The six processed conditions consisted of each of the three signal transformation schemes performed two different ways, once with the specified acoustic parameter altered as

Table 7.1: The seven conditions tested for each talker. The order of presentation of test conditions was varied for each talker.

Speaking Mode	# of sentences
conversational/normal	30
2xprocessed(same_form)/normal	30
processed(enh_form)/normal	20
processed(same_pitch)/normal	30
processed(enh_pitch)/normal	30
2xprocessed(same_env)/normal	30
processed(enh_env)/normal	30

in Chapter 5 and once with the parameter unaltered.

The processed(same_pitch) speech was obtained by passing conv/normal speech through the LPC analysis-synthesis processing without modifying F0. For the formant and envelope signal transformation schemes, however, an analogous procedure would have essentially resulted in an identity system, and no artifacts would have been introduced into conv/normal speech. Therefore, each transformation was instead performed twice, first altering the parameter as in Chapter 5 and then returning the parameter to its original value. This two-pass procedure was used to provide a rough estimate of the degradation due to artifacts; it likely resulted in a greater degradation due to artifacts than only processing the speech once and therefore represents a worst-case scenario of the effect of artifacts.

All three of the processed/unaltered conditions produced speech with the desired acoustic properties. Measurements of the average and standard deviation of F0 verified that the pitch was unaltered after LPC analysis-synthesis (see Tables B.3 and C.2 for each of the talkers). Similarly, measurements of the long-term spectra of the twice-processed(formant) speech and of the intensity envelope spectra of the twice-processed(envelope) speech confirmed that the two-pass processing procedure roughly restored these properties to their original values (see Section B.6 and C.7).

Four normal hearing listeners (all males; age range: 21 to 27 years old) were employed to evaluate the intelligibility of the speech in the presence of additive speech-shaped[39] noise. The results of each listener's hearing test is presented in

Appendix D. Each listener heard the speech of one talker in all seven conditions, and each listener was presented speech from a different talker. Listeners selected the ear that would receive the stimuli and were tested individually in one session, roughly two hours in duration. The presentation setup was the same as the one described for the normal-hearing experiment in Chapter 6, with one small difference: an SNR of 0dB was used, since scores obtained in Chapter 6 (SNR = –1.8 dB) were fairly low.

7.2.2 Results

Subject responses were scored as outlined in Chapter 6. Percent-correct key word intelligibility scores for each talker are presented in Figure 7-5. Scores for T2, T3, and T4 were fairly consistent. First, formant processing did not appear to have any significant artifacts for these talkers, since scores for the twice-processed(same_formant) condition were essentially the same as scores for unprocessed conversational speech. Unfortunately, it was also true that processed(enhanced_formant) speech was not significantly more intelligible than conversational speech. Assuming that the probability of correctly identifying key words from each talker can be represented by a binomial distribution where p is estimated to be the mean intelligibility score observed for the corresponding condition in the previous intelligibility experiment, the probability of observing such scores for processed (enhanced_formant) speech by chance was only 0.08. Therefore, the difference in intelligibility benefit of processed(enhanced_formant) speech between the two experiments is more likely a result of the different signal to noise ratios used. At an SNR of 0dB, the formants are most likely prominent enough relative to the noise that enhancement does not improve intelligibility as substantially as it does at SNR = –1.8dB. This idea is explored further in Section 7.3.1. Second, the LPC analysis-synthesis processing used for pitch modification did degrade intelligibility to some degree. This degradation can be seen from the fact that processed(same_pitch) speech was an average of 7.6 percentage points less intelligible relative to conversational speech for T2, T3, and T4. Even when taking this degradation into account, however, the processed(enhanced_pitch) speech did not provide an intelligibility benefit for these talkers. In fact, processed(enhanced_pitch)

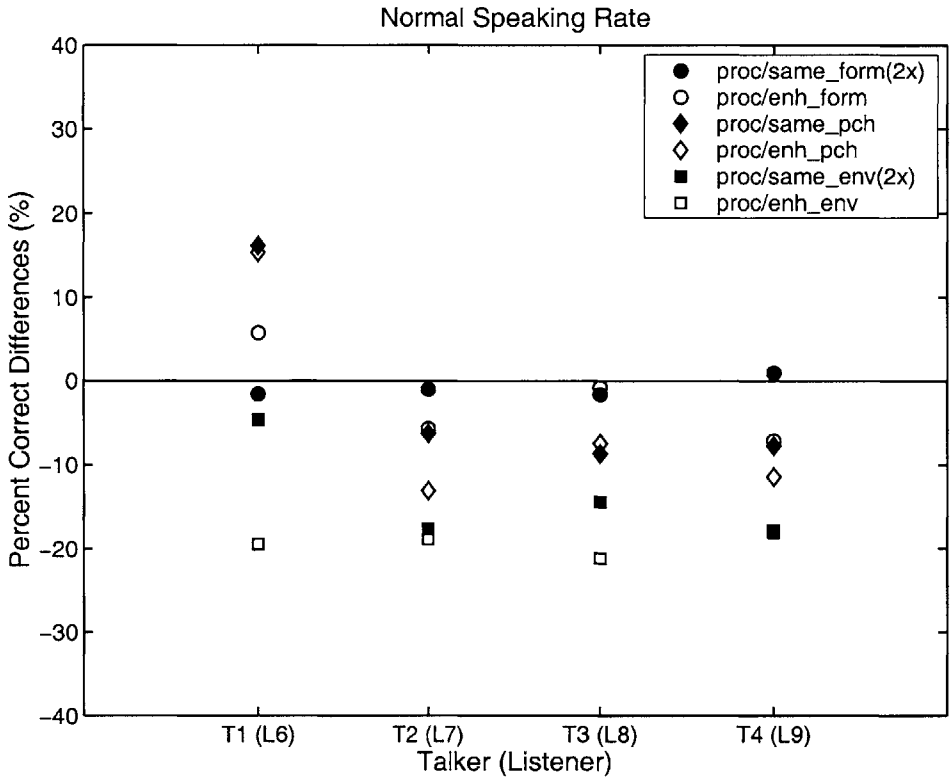

Figure 7-5: Percent correct scores relative to conversational mode for each talker (listener). Each listener heard the speech of one talker only.

speech had intelligibility near or below processed(same_pitch) speech. Finally, Figure 7-5 shows that the envelope processing scheme degraded speech intelligibility substantially (an average of 16.7 percentage points for T2, T3, and T4). Nonetheless, processed(enhanced_envelope) speech showed no benefit relative to the twice-processed(same_envelope) condition. Overall, it can be concluded that neither transforming the envelope or altering the distribution of F0 had a beneficial effect on intelligibility, even when taking processing artifacts into account. Moreover, the benefit of formant processing is limited to signal to noise ratios less than 0dB.

For T1, as with the other talkers, formant processing did not introduce any artifacts that reduced intelligibility. Moreover, unlike the other talkers, enhancement of formants provided an intelligibility benefit over conversational speech, even at a 0dB signal-to-noise ratio. Possible reasons for this result are explored further in Sec-

Table 7.2: The three conditions and corresponding 30-sentence list used to test each listener. The list tested in each condition was rotated to eliminate any chance of varying difficulty of lists effecting the outcome. The sentences from Picheny's corpus[44] used in each of these lists are specified in Appendix F.

	Listener					
Condition	L10	L11	L12	L13	L14	L15
conversational/normal	C1	C5	C4	C5	C7	C3
processed(same_pitch)/normal	C5	C4	C1	C3	C5	C7
processed(enh_pitch)/normal	C4	C1	C5	C7	C3	C5

tion 7.3.1. Although a smaller degradation in intelligibility due to envelope processing artifacts was observed (4.7 percentage points) for T1 as opposed to the other talkers, envelope enhancement still degraded intelligibility by roughly the same amount as for the other talkers. Perhaps the most unexpected result is that the LPC processing provided roughly the same amount of intelligibility benefit, regardless of whether F0 was modified or not. Since only one listener was used per talker, it was assumed that this result was spurious, due to the small sample size.

To verify this assumption, an additional intelligibility experiment involving only T1 was performed. Six normal hearing listeners (three females, three males; age range: 21 to 41 years old) participated in the experiment. The results of each listener's hearing test are listed in Appendix D. Three conditions were tested: conversational/normal, processed(same_pitch)/normal, and processed(enhanced_pitch)/normal. Although each 30-sentence list was believed to be of equal difficulty, lists were rotated so that each listener heard each list in a different condition. This procedure ensured that the difficulty of a given list would not be a factor in the results. The lists presented to each listener in each condition are specified in Table 7.2.

Percent-correct key word intelligibility scores for each listener are shown in Figure 7-6. On average, processed(same_pitch) speech was most intelligible at 63%, followed by processed(enhanced_pitch) speech at 62% and conversational speech at 56%. A t-test was applied, after an arcsine transformation ($\arcsin\sqrt{I_j/100}$) to equalize the variances, in order to determine the significance of difference between the means of each test condition. Neither of the means of the processed conditions were signifi-

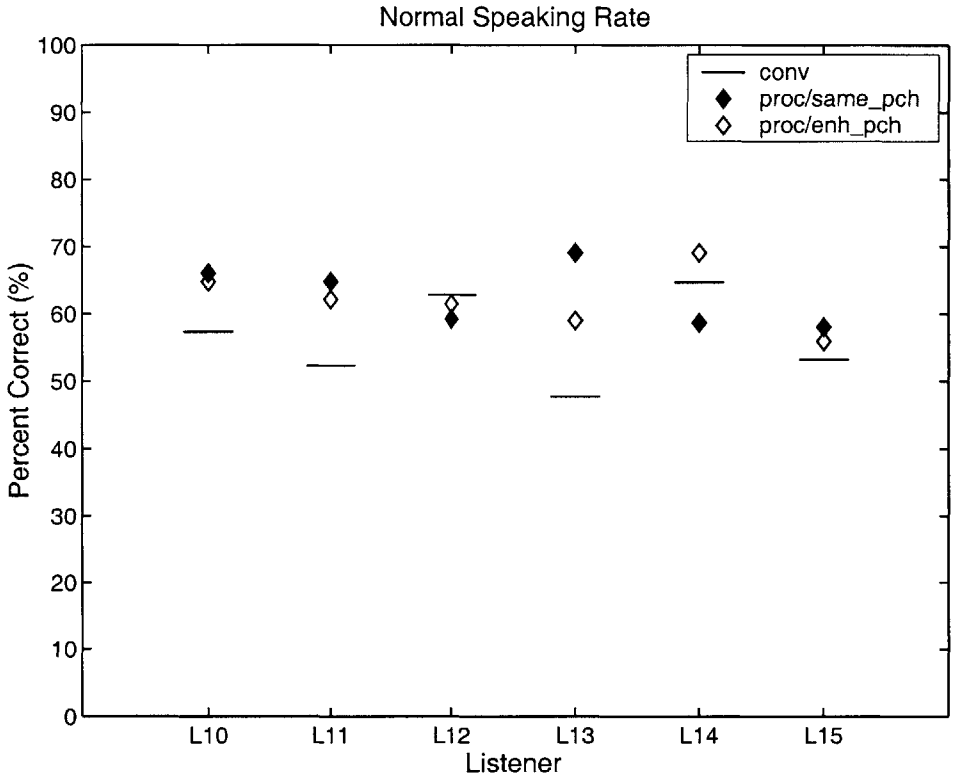

Figure 7-6: Percent correct scores relative to conversational mode for each listener.

cantly different than the conv/normal mean at the 0.05 level, after averaging scores across listener. However, the significance level for both of the processed conditions was p=0.08, which leaves open the possibility that both of these conditions provided an intelligibility benefit to listeners. Speech-based STI measurements, shown in Table 7.3, also support this possibility, predicting intelligibility improvements from both types of LPC processing (with or without pitch modification) for T1, even though intelligibility degradations from both types of processing were correctly predicted for T2.

7.3 Further Analysis of Signal Transformations

The results of the follow-up intelligibility experiment indicated that processed(C)/normal speech does not improve intelligibility, even after including the effects of process-

Table 7.3: Speech-based STI measurements for conditions tested in final intelligibility experiment with T1. STI measurements for T2 in corresponding conditions are provided for reference. These measurements support the possibility that LPC analysis-synthesis could have improved the intelligibility of T1 for normal hearing listeners in noise.

Condition	Talker	
	T1	T2
conversational/normal	0.53	0.48
processed(same_pitch)/normal	0.56	0.45
processed(enh_pitch)/normal	0.56	0.45

ing artifacts. A similar result was obtained for processed(B)/normal speech, although the LPC analysis-synthesis artifacts actually improved intelligibility for T1, a male talker. In order to obtain an explanation for this effect as well as the varying benefit of processed(A)/normal depending on signal-to-noise ratio, further analysis was conducted.

7.3.1 Formant Processing

Because processed(A)/normal speech, the transformation associated with modification of formant frequencies, provided a statistically significant intelligibility benefit when the SNR was –1.8dB but did not appear to provide a similar benefit in the follow-up experiment when the SNR was 0dB, the benefit of formant processing for normal hearing listeners seems to be associated with formant audibility. Spectrograms of sentences in noise show clearly that the processing raised the spectral prominences for the formants of T1, T2, and T3 to a level above the noise. A typical spectrogram for these talkers is shown in Figure 7-7. The reason the intelligibility benefit was not as significant for T4 at SNR=–1.8dB probably stems from the fact that a high percentage of her formants in conv/normal speech were already above the level of the noise, as shown in Figure 7-8. Similarly, when the experiment was repeated at an SNR of 0dB, the percentage of audibile formants was improved relative to the previous experiment (SNR=–1.8dB) for all talkers. Therefore, raising the level of the formants at this SNR could not provide as much improvement to formant audibility

as could be achieved at an SNR 0f −1.8dB.

7.3.2 Pitch Processing

After accounting for the effect of artifacts, processed(B)/normal speech, the transformation associated with pitch modification, did not provide an intelligibility benefit. However, perhaps because LPC analysis-synthesis provided a high-frequency boost, some intelligibility improvement was observed for T1. From examination of long-term average spectra, shown in Figures 7-9 through 7-12, it appears that the reason this high-frequency emphasis improved the intelligibility for T1 only is because the emphasis was largest in magnitude and extended into higher frequencies for this talker, thus raising the levels of his third and fourth formants above the level of the noise. Inspection of individual vowel spectra, shown in Figure 7-13, also suggests that for T1, LPC analysis-synthesis may have increased the peak levels of the second through fourth formants relative to the level of the first formant, an effect similar to that achieved with formant processing. If this were the case for a high percentage of T1's vowels, the intelligibility benefit observed for T1's processed(B)/normal speech was simply a replication of the advantage observed for processed(A)/normal speech. However, it is not known why LPC processing would have had such an effect on any talker or on T1 in particular. Thus, it is possible that the formant enhancement observed in T1's vowel spectra was simply caused by the high-frequency boost observed for all talkers. In this case, another possible factor contributing to the observed intelligibility differences could be that the quality of LPC synthesis for the other talkers may not have been as high as for T1, and degradations due to those artifacts counteracted the benefit of the high-frequency emphasis.

Although high-frequency emphasis is similar to the idea behind modifying formant frequencies, formant frequency modification was much more successful in improving intelligibility for all talkers except T1. Because the formant processing is concentrated only in the F2 and F3 region, a larger emphasis was provided to these frequencies than by the high-frequency boost associated with LPC analysis-synthesis. Figures 7-9 through 7-12 clearly show this effect. From these figures, it appears that formants

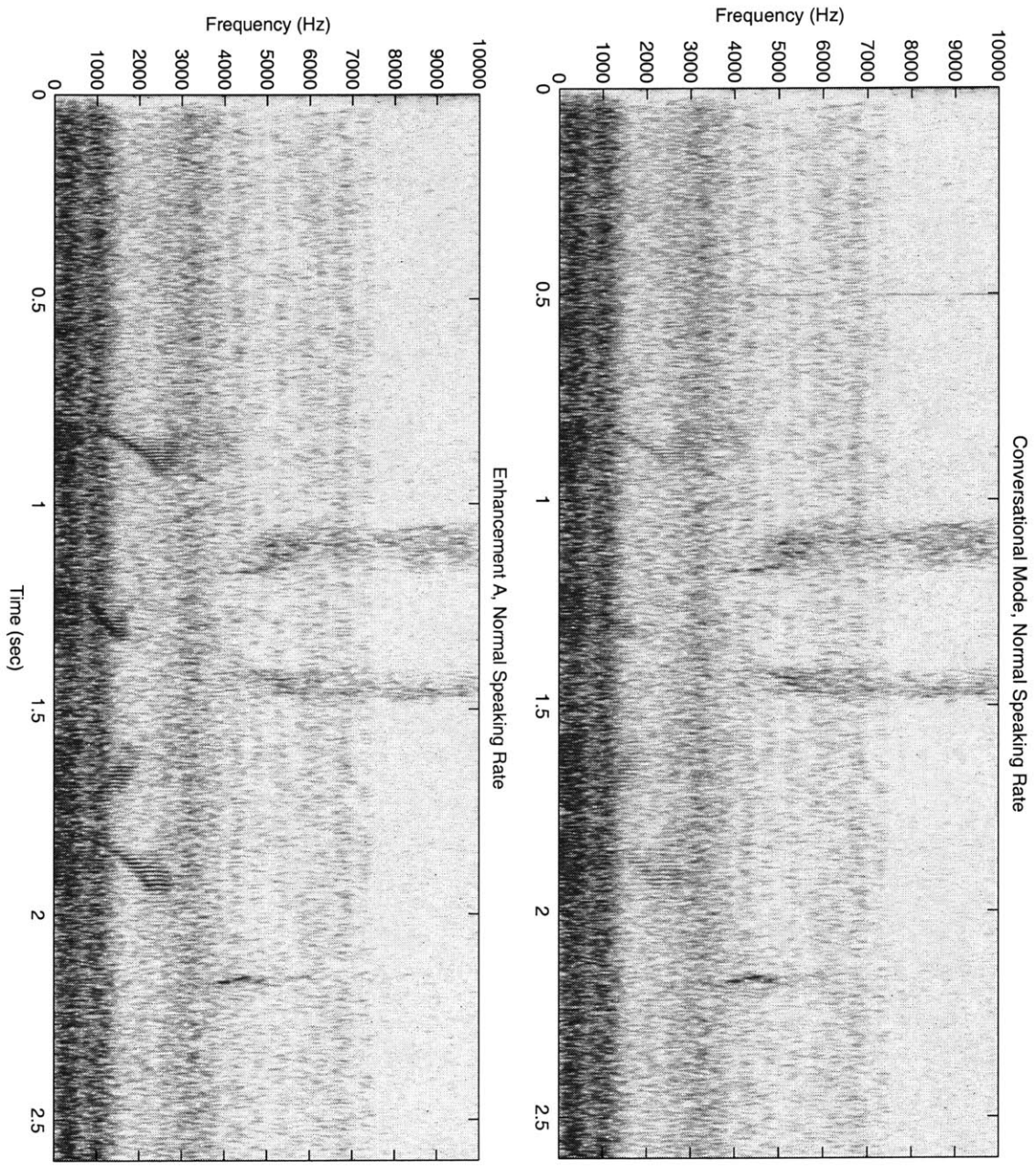

Figure 7-7: Spectrograms for T1's sentence, "A plate sorts their wait," before and after processing(A), at SNR=–1.8dB. After processing, spectral prominences associated with F2 and F3 were more frequently above the level of the noise, resulting in improved intelligibility.

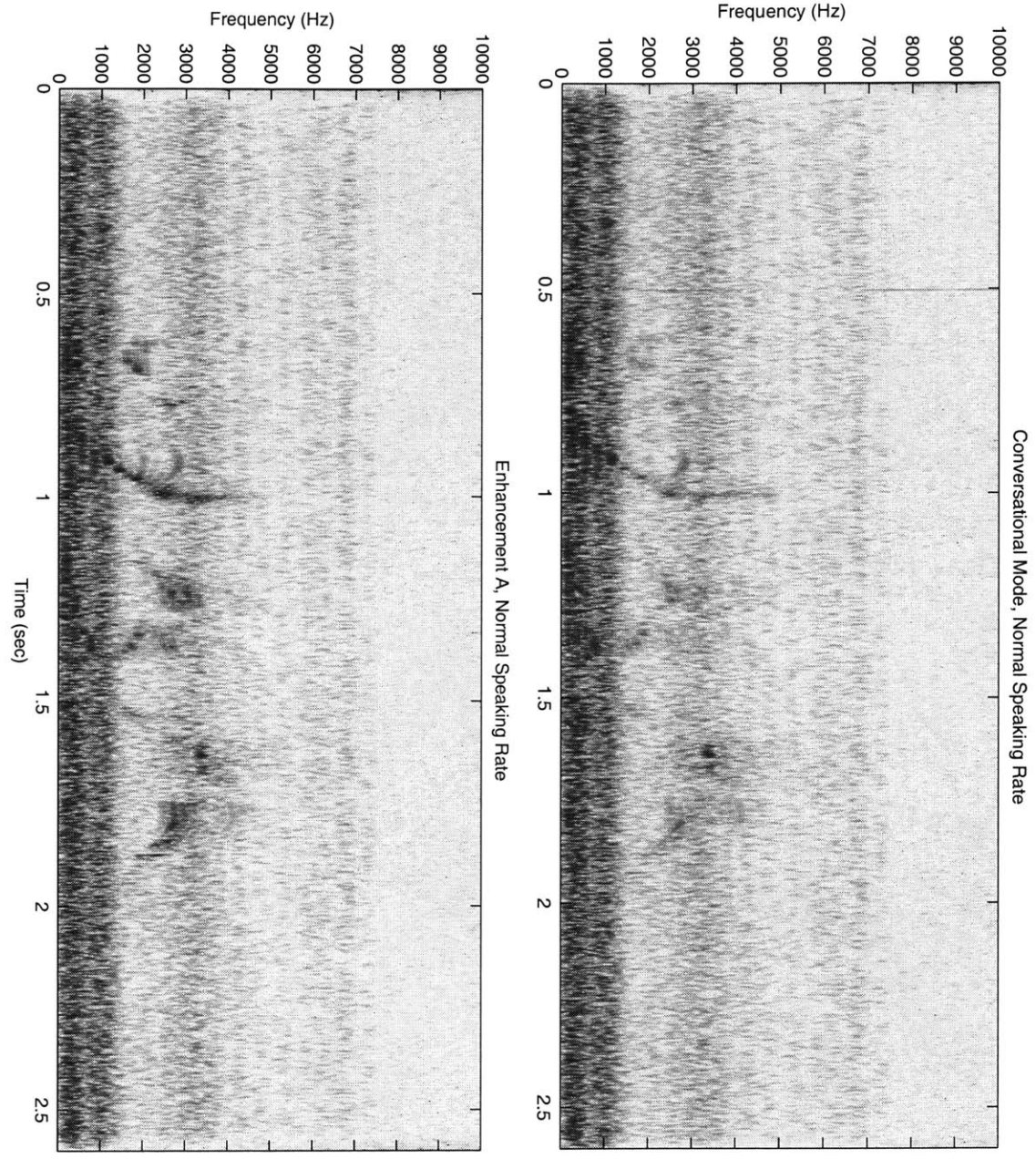

Figure 7-8: Spectrograms for T4's sentence, "Your joint made up a shear," before and after processing(A), at SNR=–1.8dB. Before processing, spectral prominences associated with F2 and F3 were often above the level of the noise. Therefore, processing could not substantially increase formant audibility and no significant intelligibility improvement was observed.

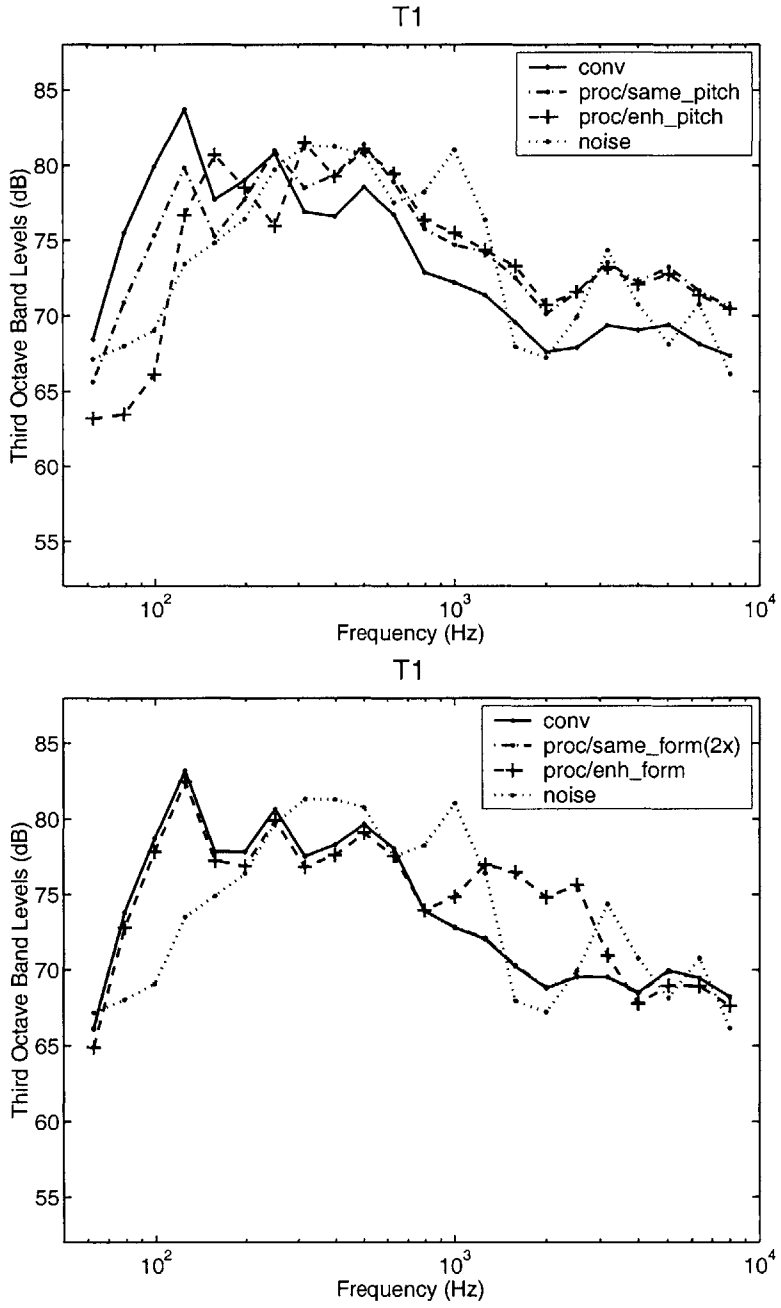

Figure 7-9: Third-octave band levels for T1's conversational, processed(same_pitch), and processed(enhanced_pitch) speech at normal rates. Third-octave band levels for the speech-shaped noise is also provided for reference.

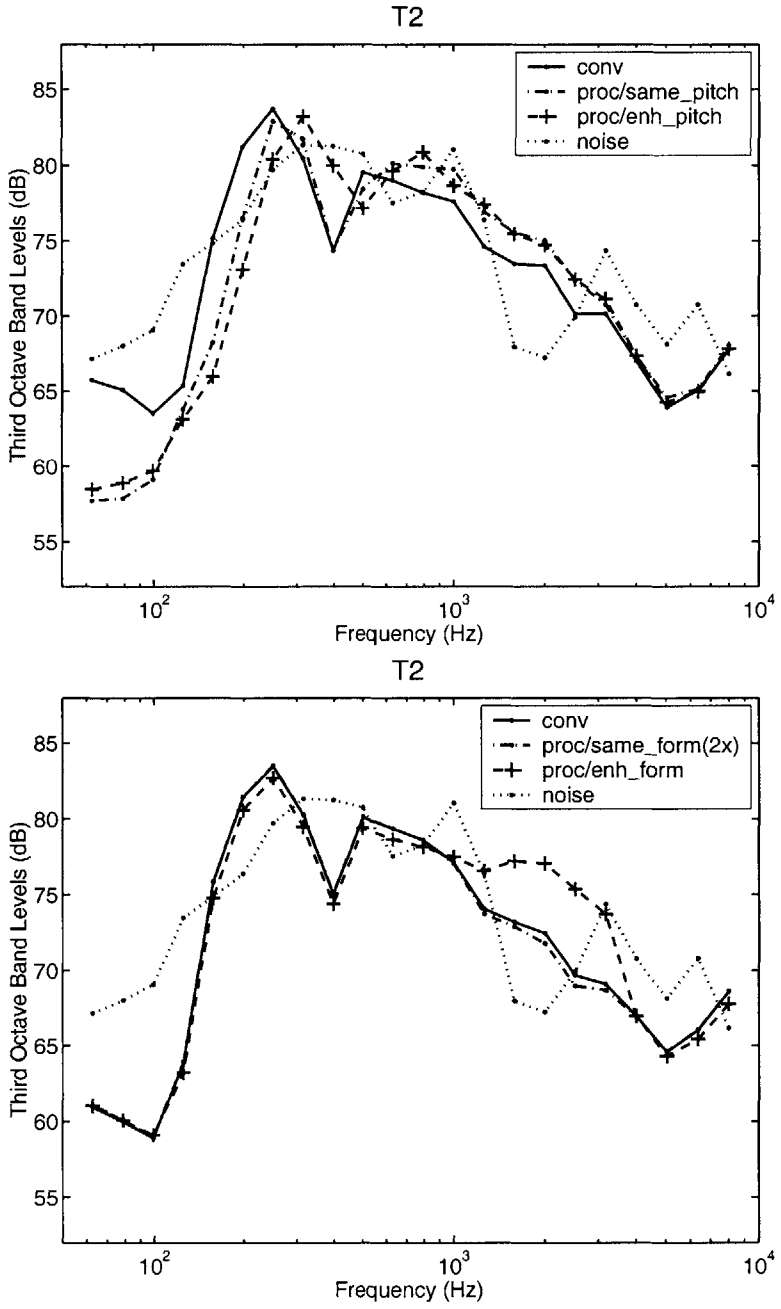

Figure 7-10: Third-octave band levels for T2's conversational, processed(same_pitch), and processed(enhanced_pitch) speech at normal rates. Third-octave band levels for the speech-shaped noise is also provided for reference.

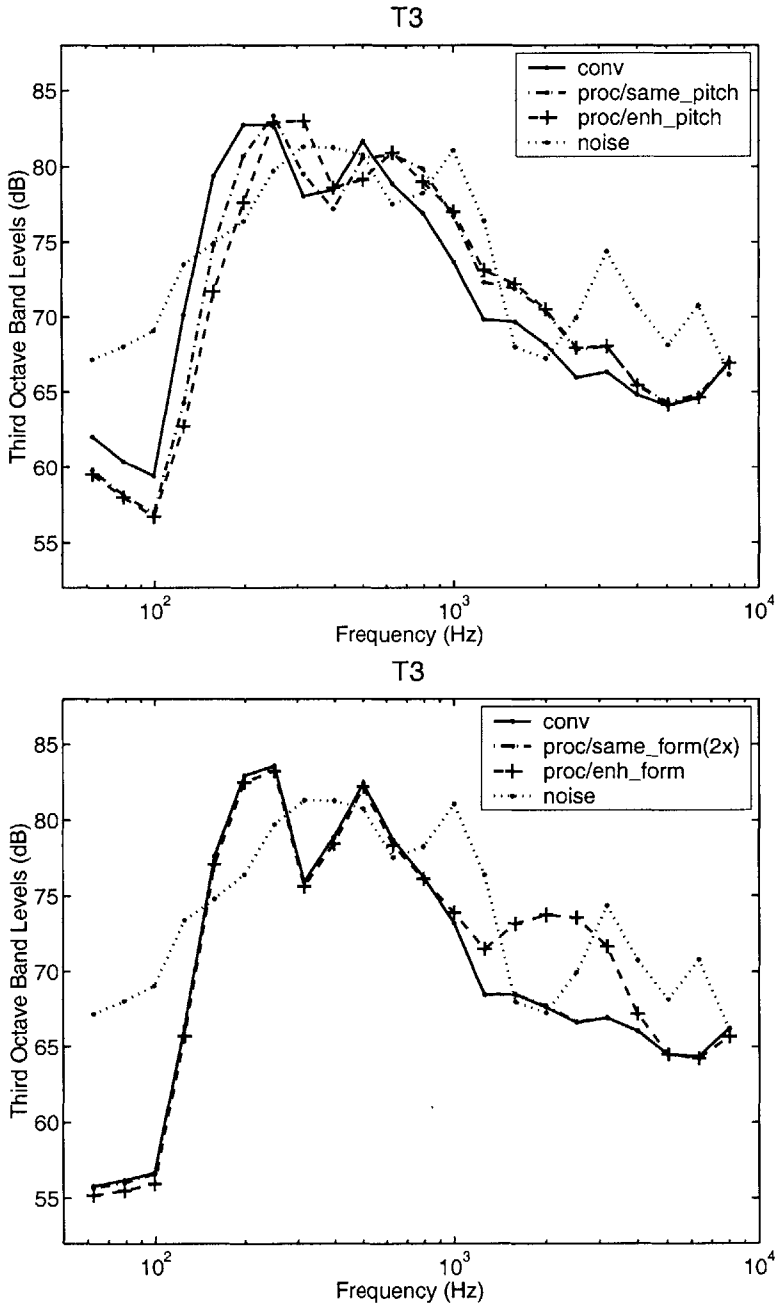

Figure 7-11: Third-octave band levels for T3's conversational, processed(same_pitch), and processed(enhanced_pitch) speech at normal rates. Third-octave band levels for the speech-shaped noise is also provided for reference.

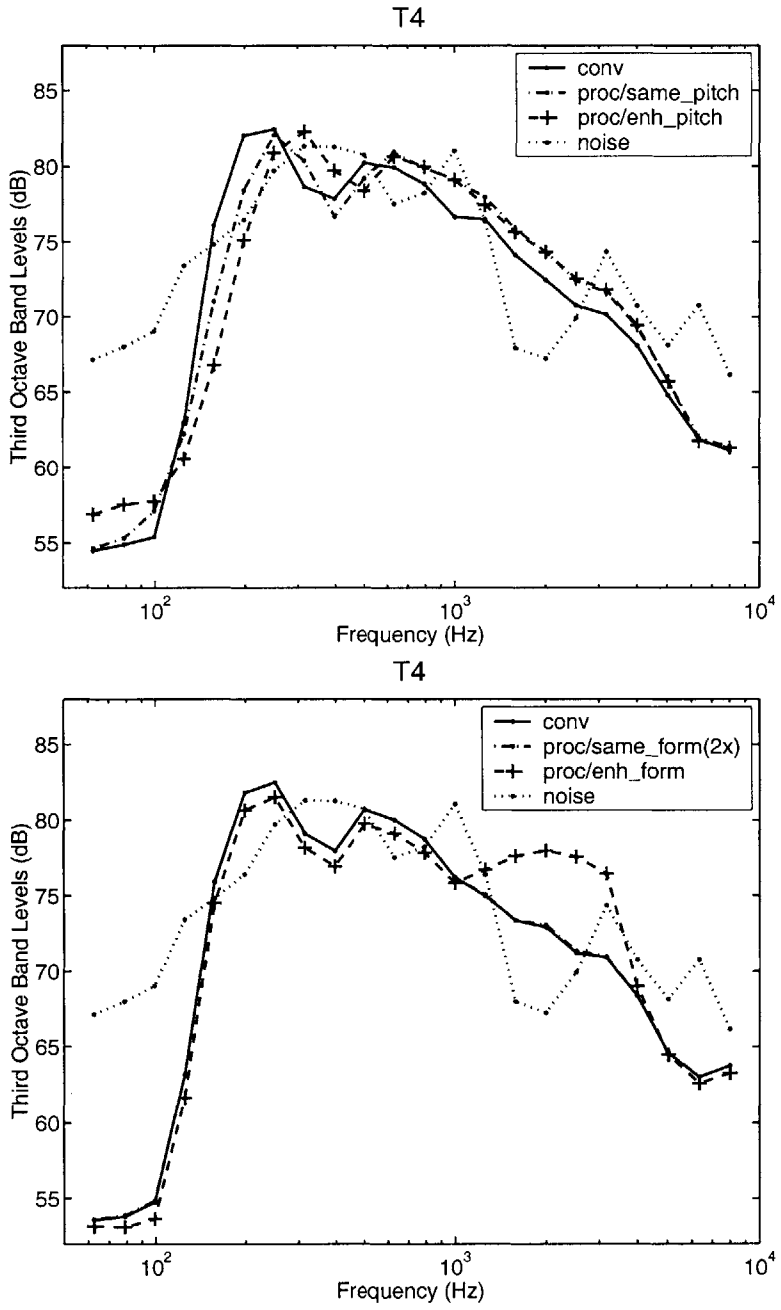

Figure 7-12: Third-octave band levels for T4's conversational, processed(same_pitch), and processed(enhanced_pitch) speech at normal rates. Third-octave band levels for the speech-shaped noise is also provided for reference.

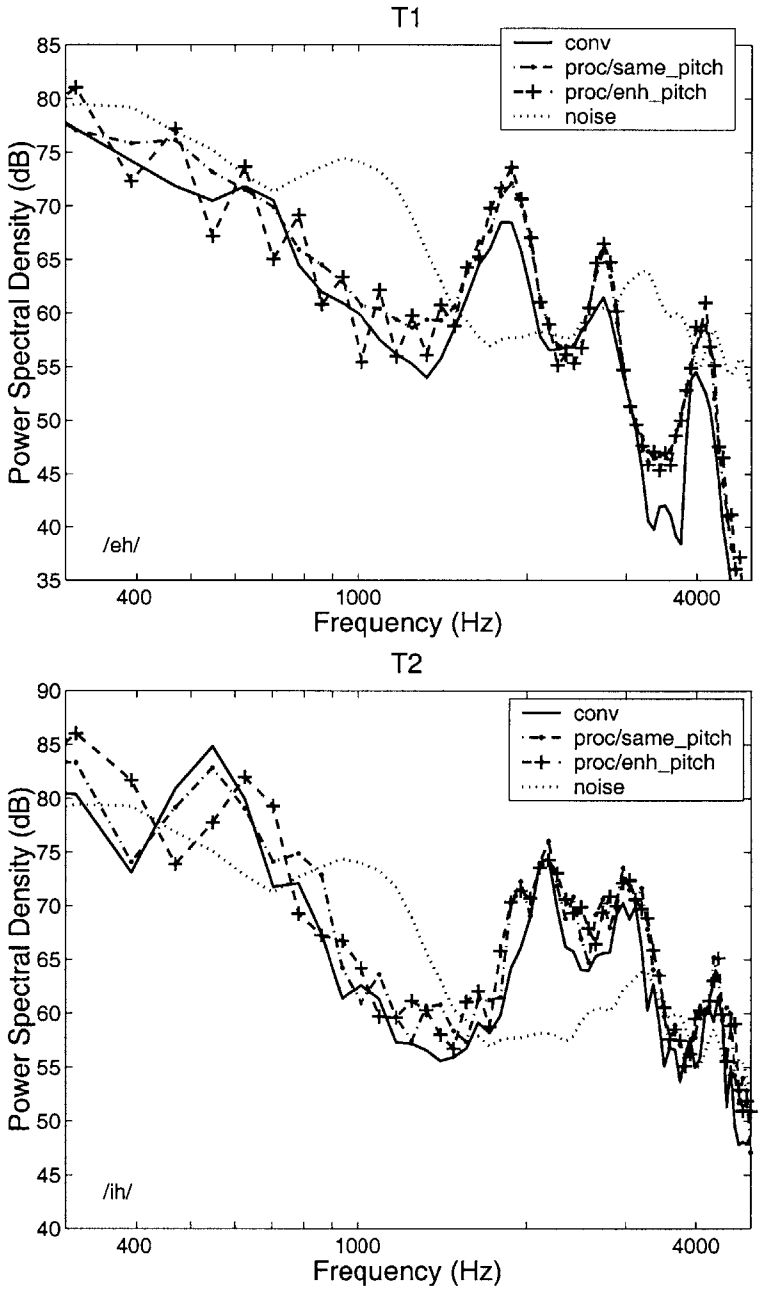

Figure 7-13: Power spectral density of an individual vowel token in conversational, processed(same_pitch), and processed(enhanced_pitch) speech at normal rates. Power spectral density of noise is also provided for reference. LPC analysis-synthesis appears to have increased the level of T1's second through fourth formants relative to his first formant but no such effect was observed for T2.

were audible over the noise for all talkers after formant processing but only for T1 after LPC analysis-synthesis.

7.4 Summary

Of the processing schemes examined in this study, only formant enhancement appears to have improved intelligibility, after accounting for the effects of processing artifacts. Moreover, the benefits of this type of processing appear to be linked primarily to formant audibility. The follow-up intelligibility experiment showed that the benefit of processed(B)/normal speech for T1 was not associated with the modification of F0 but only with the effects of LPC processing. Additional acoustic measurements showed that these effects included some modification of both the envelope spectra and the long-term spectra. The intelligibility experiment also showed that envelope enhancement did not provide an intelligibility benefit, even after accounting for the effects of processing artifacts.

Chapter 8

Conclusion

This thesis examined acoustic properties of clear speech at normal speaking rates. Three global acoustic properties associated with clear/normal speech were identified: increased energy near the second and third formants, higher average and greater range of F0, and increased modulation depth of low frequency modulations of the intensity envelope. Signal processing schemes that altered these properties in conv/normal speech to approximate their characteristics in clear/normal speech were then developed. Results of intelligibility tests, however, suggest that these properties may not fully account for the intelligibility benefit of clear/normal speech. Other properties important for highly intelligible speech may have been difficult to measure due to the complexity of the acoustic database (see below) and varying talker strategies.

8.1 Intelligibility Results

Although previous clear speech studies found intelligibility results for hearing-impaired listeners to be consistent with results for normal hearing listeners in noise[42, 55], such consistency across populations was not observed in this study. In fact, of the two conditions that provided an intelligibility benefit over conv/normal speech for normal hearing listeners (clear/normal speech and processed(A)/normal speech), neither provided a statistically significant benefit to hearing-impaired listeners, when averaged across talker and listener. One possibility is that the intelligibility

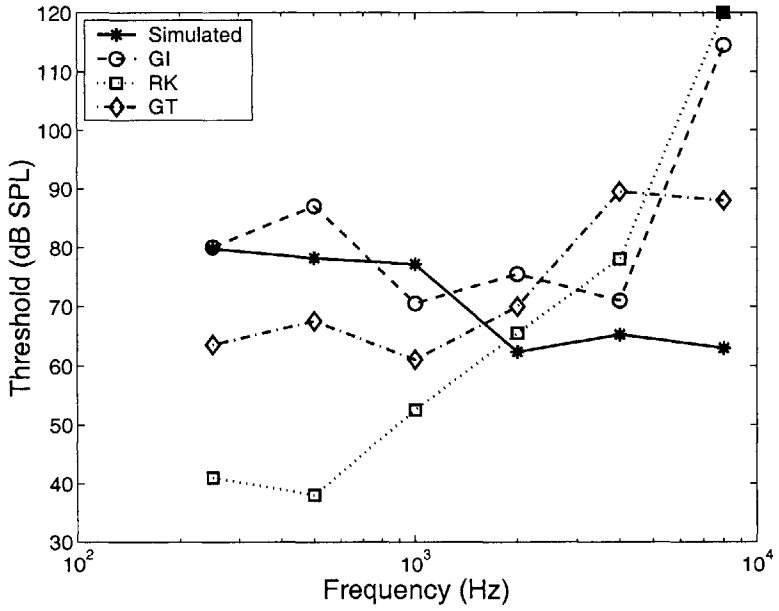

Figure 8-1: Auditory detection thresholds (dB SPL) for hearing-impaired listeners after NAL prescription and overall level adjustment (each listener selected a comfortable listening level). Filled square indicates detection threshold above 120 dB SPL. Solid line represents the impairment simulated in normal hearing listeners by the addition of noise.

benefits of these conditions do not extend to hearing-impaired listeners and that the additive noise model for simulating impairment in normal hearing listeners is inadequate. To examine this possibility more closely, the hearing loss simulated in this study was determined from masking levels of pure tones resulting from the additive noise, calculated from critical ratios[19]. Figure 8-1 compares the elevated thresholds experienced by normal hearing listeners as a result of the additive noise with the auditory detection thresholds of each hearing-impaired listener, after accounting for the NAL prescription and overall gain adjustment used in the experiment. Clearly, the simulated impairment was only a crude approximation of the actual listeners' hearing losses. Therefore, it is possible that the intelligibility benefit of clear/normal speech experienced by the normal hearing listeners with simulated impairment may not extend to hearing-impaired listeners.

Another possibility is that the benefits of clear/normal speech may be related to

age, since the hearing-impaired participants in this study were older (40 to 65 years) than the normal hearing participants (19 to 43 years). A final possibility is that these conditions do provide some benefit for hearing-impaired listeners, but it was not statistically significant in this study. There is some evidence for the latter possibility, since the intelligibility experiment with hearing-impaired listeners showed both that clear/normal speech was 7 percentage points more intelligible than conv/normal speech on average and that T3's processed(A)/normal speech was significantly more intelligible than conv/normal speech.

In addition, the hearing-impaired intelligibility data suggest that an individual's audiometric characteristics may also be a factor in whether clear/normal speech can be of benefit. A close inspection of interactions between hearing-impaired listener and talker reveals that RK received little or no benefit from clear/normal speech, except when listening to T2's speech, while each of the other two listeners experienced moderate to large intelligibility gains from clear/normal speech for all talkers. Since RK also had the most precipitous hearing loss, it is possible that listeners with this type of audiometric characteristic may not benefit from clear/normal speech, and the additive noise model is inappropriate for simulating such hearing losses. This possibility is supported by Figure 8-1, which shows that the elevated detection thresholds simulated for normal hearing listeners differ more from RK's thresholds than from the thresholds of the other hearing-impaired listeners. To address the question of whether audiometric characteristics are linked to an individual's ability to benefit from clear/normal speech, additional intelligibility tests targeting groups of listeners with various types of audiometric characteristics would be required.

For normal hearing listeners, both clear/normal and processed(A)/normal speech provided an intelligibility benefit over conv/normal speech. However, the benefit afforded by processed(A)/normal speech (6 percentage points on average) was less than half that afforded by clear/normal speech (16 percentage points), suggesting that increased energy near second formants is not the only factor responsible for the intelligibility advantage of clear/normal speech. Moreover, the follow-up intelligibility experiment demonstrated that neither processed(B)/normal speech nor pro-

cessed(C)/normal speech provided an intelligibility benefit over processed/unaltered speech containing similar processing artifacts. Because these three properties, alone or in combination, did not reproduce the large intelligibility gains observed for clear/normal speech, additional acoustic properties that were not identified in this study must also contribute to the high intelligibility of clear/normal speech.

8.2 Acoustic Database

One factor that may have prevented identification of one or more additional properties of highly intelligible speech is the complexity of the acoustic database. In Krause's original study[28], the sentence database was well-suited for the task of eliciting clear/normal speech from talkers, providing talkers with a natural, familiar task and the opportunity to manipulate the acoustics of an entire sentence instinctively, including global and phonological variables as well as phonetic variables. For the purposes of acoustic analysis, however, a database of sentences poses some difficulties.

The primary problem with using a sentence database for the purposes of acoustic analysis is the presence of acoustic variability due to word positioning within sentences or phonetic context within words. For some acoustic properties, this variability could be large enough to mask the variability between conv/normal and clear/normal tokens. A typical example of the number of environments in which a phone occurs in the database used for this study is the phone, /eh/. Of the 36 occurrences of /eh/ in the list of 50 sentences used for acoustic analysis of SA, eleven occurrences were in function words and twenty-five in content words. Furthermore, the 25 occurrences of /eh/ in content words appeared in 22 unique phonetic environments, if the phonetic context of both the preceding and subsequent phone is considered. Thus, when the tokens were analyzed as a group, token variability due to environment may have been large enough to mask variability due to speaking mode, even when function words and content words were analyzed separately. However, with so few tokens in each phonetic environment, a meaningful statistical analysis of speaking mode differences within specific environments could not be performed. Therefore, it seems possible that

a database with only a fixed set of environments and many tokens in each environment could have uncovered acoustic properties associated with clear/normal speech in addition to those identified in this study from the sentence database.

8.3 Talker Strategies

A second factor which affected the outcome of the acoustic analysis is varying talker strategies. While SA and RG both exhibited acoustic properties of clear/slow speech that were consistent with those previously reported[6, 44], each retained a somewhat different subset of those acoustic properties when the constraint of maintaining normal speaking rate was imposed. In addition, the intelligibility advantage of SA's clear/normal speech was more robust to other degradations (high-pass filtering, reverberation, non-native listeners) than RG's clear/normal speech. Thus, the two talkers appear to have implemented different strategies for producing clear speech at normal rates.

Given that SA and RG used differing strategies, it is reasonable to assume that other talkers in Krause's original study[28] may also have employed unique strategies for producing clear/normal speech. If this is the case, the results of the acoustic analysis in this study were very much dependent on selection of talkers. If T3 and T4 had been analyzed rather than RG and SA, a different set of acoustic properties associated with clear/normal speech might have been identified. Thus, additional insights into the properties of clear/normal speech may be obtained by acoustic analysis of these and other talkers trained to produce clear speech at normal rates.

Although this study identified different talker strategies for producing clear/normal speech, one similarity between the talkers was also observed. Both talkers omitted a number of acoustic properties in clear/normal speech that were present in clear/slow speech. These omissions were most most likely caused by physiological constraints on articulation at normal speaking rates that hindered the simultaneous expression of such a large number of acoustic properties. The properties that were omitted, however, varied for each talker, and the reasons for this variation are not known. Possibilities include the degree of articulation effort and the perceived intelligibility

benefit associated with producing a given acoustic characteristic in a particular environment. It is even possible that each talker's strategy represents a tradeoff between articulation effort and increased intelligibility, optimized according to his/her unique vocal characteristics and the acoustic characteristics of the surrounding environment. If so, processing schemes based primarily on the acoustics of one talker's clear/normal speech, such as those developed in this study (based primarily on the acoustics of SA's robust clear/normal speech) may not be as effective for improving intelligibility of other talkers. Instead, processing schemes would need either to be tailored for specific talkers in specific environments or to encompass all talker strategies.

8.4 Suggestions for Future Work

Intelligibility experiments in this thesis included hearing-impaired listeners ranging in age from 40 to 65 years old and relatively young (19 to 43 years old) adult listeners with normal hearing, using additive noise to simulate hearing loss. Although this simulation is appropriate for many mild to moderate impairments, it may not represent the effects of more severe impairments accurately. Moreover, some studies report an age-related decline in speech reception for elderly listeners[1], particularly those with hearing impairments[18]. Therefore, additional intelligibility tests should be conducted to evaluate the intelligibility of clear/normal, clear/slow, conv/normal, and conv/slow speech for young hearing-impaired, elderly hearing-impaired, and elderly normal-hearing listeners. These tests would differentiate the effect of age and impairment factors and clearly identify which groups can receive benefit from clear speech at normal speaking rates.

Of the processing schemes examined in this study, processed(A)/normal speech provided the only consistent intelligibility advantage over conv/normal speech for normal hearing listeners in noise. Moreover, this condition improved the intelligibility of one talker for hearing-impaired listeners, the only statistically significant intelligibility advantage observed for these listeners. However, the effect of this processing scheme was similar to high-frequency emphasis of the speech spectrum, a

frequency-gain characteristic commonly found in conventional hearing aids. Because the processing affected vowels and other voiced segments only, the increase in level of F2 and F3 relative to F1 was somewhat greater than what would have resulted by applying a high-frequency emphasis to the entire sentence, since the long-term RMS level of each sentence was normalized. Whether this additional boost to F2 and F3 provides an intelligibility benefit beyond that provided by high-frequency emphasis should be explored in additional intelligibility tests that compare the effects of high-frequency emphasis to processed(A)/normal speech and conv/normal speech. In addition, for T1, the intelligibility benefits of high-frequency emphasis should be compared to the possible intelligibility benefit for normal hearing listeners in noise of LPC analysis-synthesis of this talker. If the intelligibility benefit of LPC analysis-synthesis is replicated and cannot be attributed entirely to the effects of high-frequency emphasis, further acoustic analyses of the effects of LPC processing on T1's speech is warranted.

Although the intelligibility advantage of LPC processing for T1 as well as the intelligibility benefits of clear/normal and processed(A)/normal speech were predicted by the speech-based STI, this measure was not able to identify the relative intelligibility of all conditions correctly. In particular, several conditions that degraded intelligibility relative to conv/normal speech were predicted by the STI to provide an intelligibility benefit. Examining the acoustic characteristics of these conditions could lead to modifications of the speech-based STI that would improve the accuracy of its predictions for different styles of speech.

For the purposes of conducting further acoustic analyses and intelligibility tests of clear/normal speech, a new database of conv/normal and clear/normal speech should be created. The speech could be elicited using techniques similar to those described in Krause's original study[28]. In order to reduce acoustic variability due to context, the database should consist of a restricted set of phonetic environments, such as a database of nonsense syllables similar to that used by Chen[6]. However, it will also be necessary to use the database for intelligibility experiments, and sentences would provide richer intelligibility information than nonsense syllables, allowing for resolution

of more subtle intelligibility differences. Moreover, sentence-level speaking rates can be controlled with Krause's elicitation methods[28], but it is not known whether these methods would be as effective in controlling syllable-level rates. Therefore, sentences with a fixed number of phonetic contexts would best satisfy the conflicting demands of acoustic analyses and intelligibility experiments. It is recommended that at least 300 sentences, consisting of 3-4 key words, be constructed for each talker. If possible, the same key words should then be rearranged to create unique sentence sets for each of the other talkers. This type of corpus would allow not only for acoustic comparisons across talker but also intelligibility tests of multiple talkers, uncompromised by sentence repetitions. After construction of such a sentence corpus, a large number of talkers should be recorded so that the database will be representative of various talker strategies. An acoustic analysis on a database of this type is likely not only to identify additional acoustic properties associated with clear/normal speech but also to provide a comprehensive description of a variety of talker strategies. This information will be essential to the development of processing schemes that can provide robust intelligibility improvement for a variety of talkers and environments.

Appendix A

Acoustics Data

A.1 Global Measurements

This section contains results of global measurements, described in Section 3.2.

Table A.1: Pause length distributions.

Mode	RG		SA	
	Mean (ms)	St Dev (ms)	Mean (ms)	St Dev (ms)
Conv/Norm	41	0.5	59	0.6
Clr/Slow	487	25.5	237	30.5
Clr/Norm	76	1.8	61	1.2

Table A.2: Fundamental frequency distributions.

Mode	RG		SA	
	Mean (Hz)	St Dev (Hz)	Mean (Hz)	St Dev (Hz)
Conv/Norm	216	49	105	18
Clr/Slow	196	49	154	36
Clr/Norm	209	40	138	25

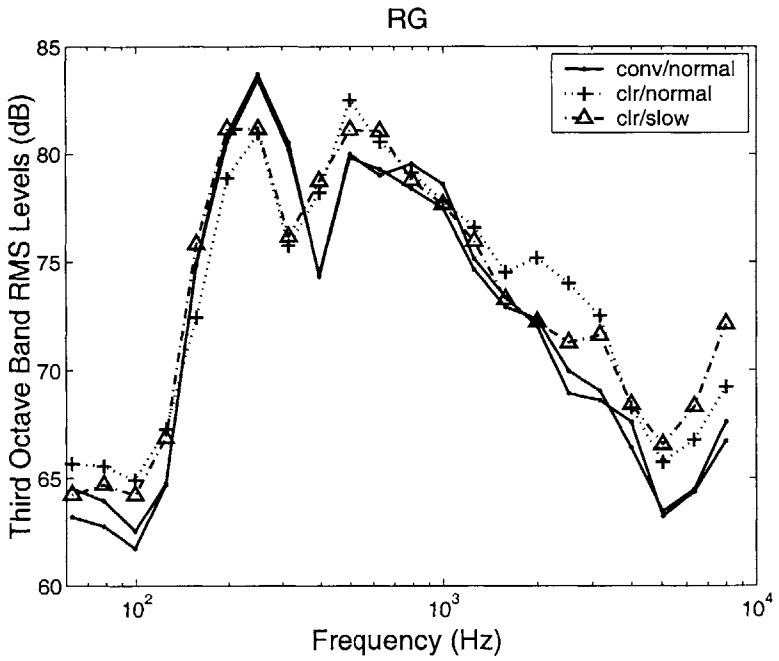

Figure A-1: Third-octave band RMS spectral levels for RG's conv/normal spectrum, clear/normal and clear/slow spectra.

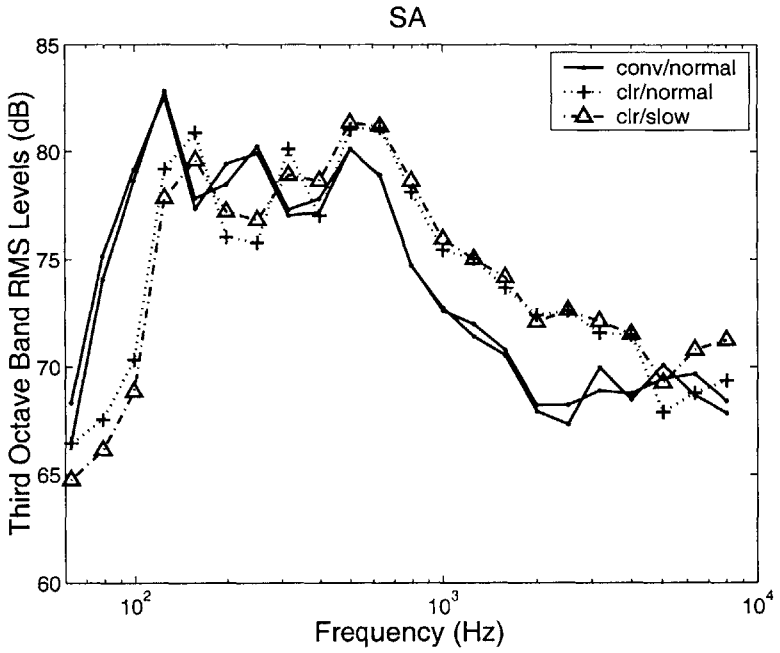

Figure A-2: Third-octave band RMS spectral levels for SA's conv/normal, clear/normal and clear/slow spectra.

A.2 Phonetic Measurements

This section contains results of phonetic measurements, described in Section 3.4.

Table A.3: Key-word formant frequency variance data (means in Hz) for RG in conversational and clear modes at normal rate. Table shows only cases ("N" represents total number of cases) where there was a significant difference in paired t-tests at the alpha=0.05 level.

Phone	Formant	StDev Diff	Conv Mean	Conv StDev	Clear Mean	Clear StDev	N	Corr	Alpha
AA	F3	-98	2672	260	2584	358	19	0.844	0.011
AO	F1	152	790	280	651	128	6	0.771	0.027
AY	F3	-164	2753	115	2727	279	8	0.707	0.006
ER	F1	52	494	116	514	64	11	0.365	0.038
UW	F1	-93	342	72	422	165	6	0.909	0.006

Table A.4: Key-word formant frequency variance data (means in Hz) for SA in conversational and clear modes at normal rate. Table shows only cases ("N" represents total number of cases) where there was a significant difference in paired t-tests at the alpha=0.05 level.

Phone	Formant	StDev Diff	Conv Mean	Conv StDev	Clear Mean	Clear StDev	N	Corr	Alpha
AH	F2	-268	1282	276	1506	544	20	0.223	0.003
AO	F2	103	1104	238	1021	135	4	0.983	0.022
AY	F1	32	706	59	729	27	8	0.628	0.016
EH	F1	44	541	113	561	69	24	0.636	0.002
EH	F2	-44	1731	133	1786	221	24	0.906	0.000
EY	F1	36	490	100	524	64	25	0.645	0.004
EY	F2	-62	1873	144	1916	206	25	0.638	0.016

Table A.5: Word-initial formant frequency variance data (means in Hz) for RG in conversational and clear modes at normal rate. Table shows only cases ("N" represents total number of cases) where there was a significant difference in paired t-tests at the alpha=0.05 level.

Phone	Formant	StDev Diff	Conv Mean	Conv StDev	Clear Mean	Clear StDev	N	Corr	Alpha
AH	F2	-162	1454	248	1673	410	22	0.391	0.010
AX	F3	-96	2915	126	2985	222	8	0.957	0.001
EH	F2	-52	1959	103	1893	155	15	0.845	0.007
EH	F3	58	2850	254	2811	196	15	0.863	0.043
ER	F1	30	384	101	478	71	24	0.213	0.049
ER	F3	81	1902	234	1796	153	24	0.622	0.008
IY	F1	-32	320	18	324	50	5	0.944	0.004
UH	F3	-70	2652	77	2495	147	7	0.777	0.028

Table A.6: Word-initial formant frequency variance data (means in Hz) for SA in conversational and clear modes at normal rate. Table shows only cases ("N" represents total number of cases) where there was a significant difference in paired t-tests at the alpha=0.05 level.

Phone	Formant	StDev Diff	Conv Mean	Conv StDev	Clear Mean	Clear StDev	N	Corr	Alpha
AO	F2	-108	1109	161	1193	269	13	0.309	0.044
EH	F1	24	518	88	519	64	21	0.767	0.019
EH	F2	-97	1697	114	1747	211	21	0.927	0.000
EY	F1	36	485	95	517	59	18	0.536	0.015
EY	F2	-89	1868	113	1964	202	18	0.778	0.001
IH	F3	63	2670	228	2685	165	28	0.236	0.048
UW	F3	50	2466	83	2613	33	3	0.988	0.047

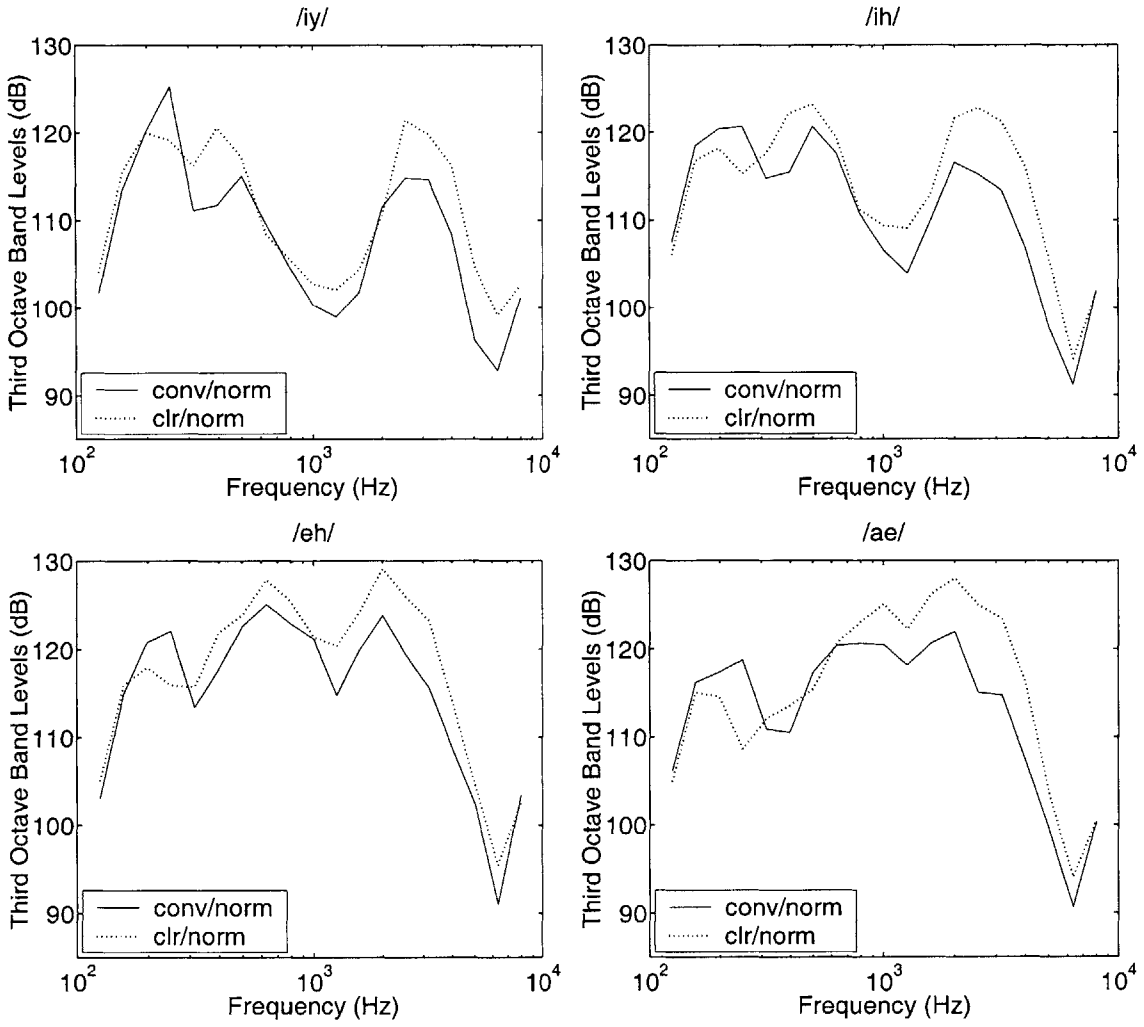

Figure A-3: Third-octave band spectra of high vowels in conv/normal and clear/normal modes for RG.

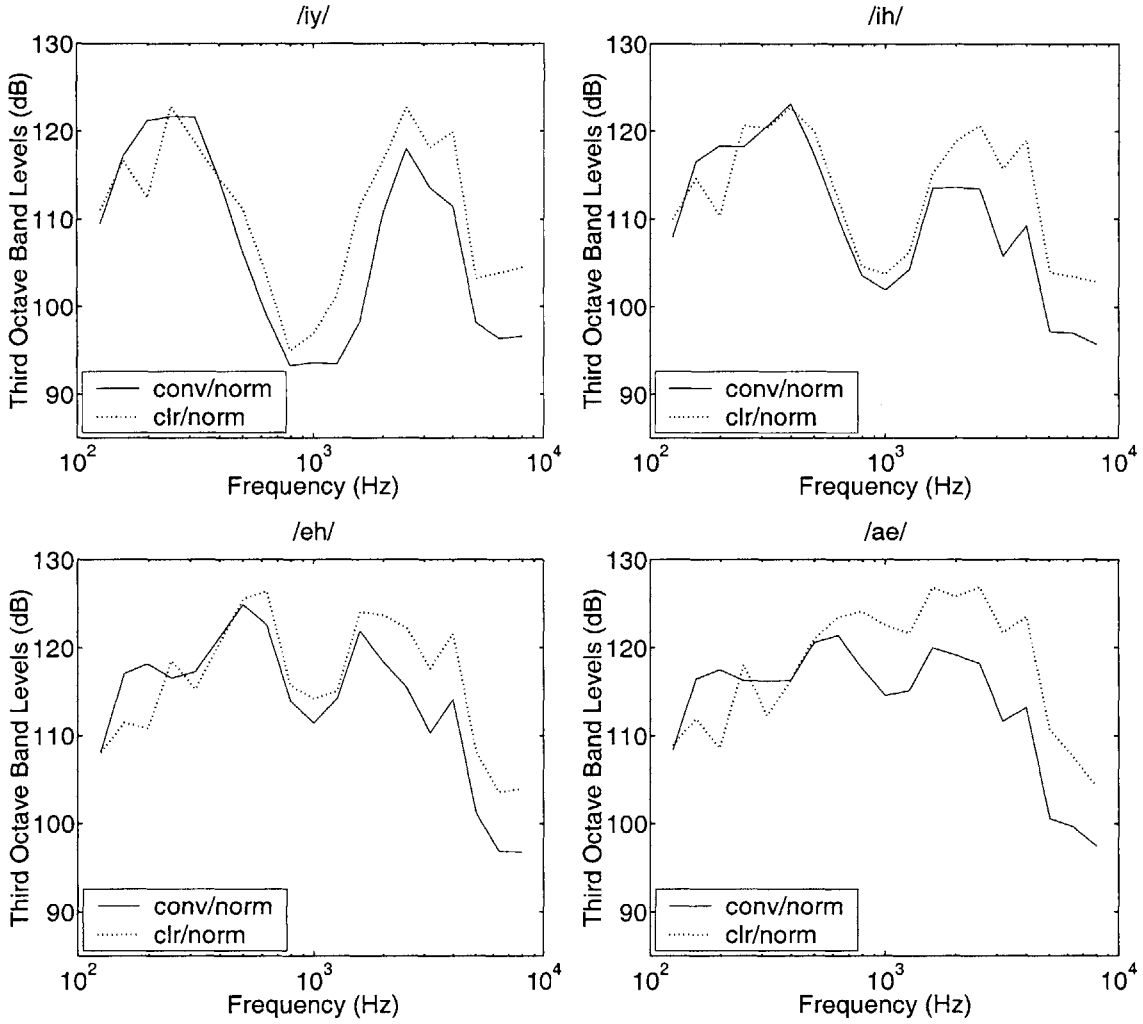

Figure A-4: Third-octave band spectra of high vowels in conv/normal and clear/normal modes for SA.

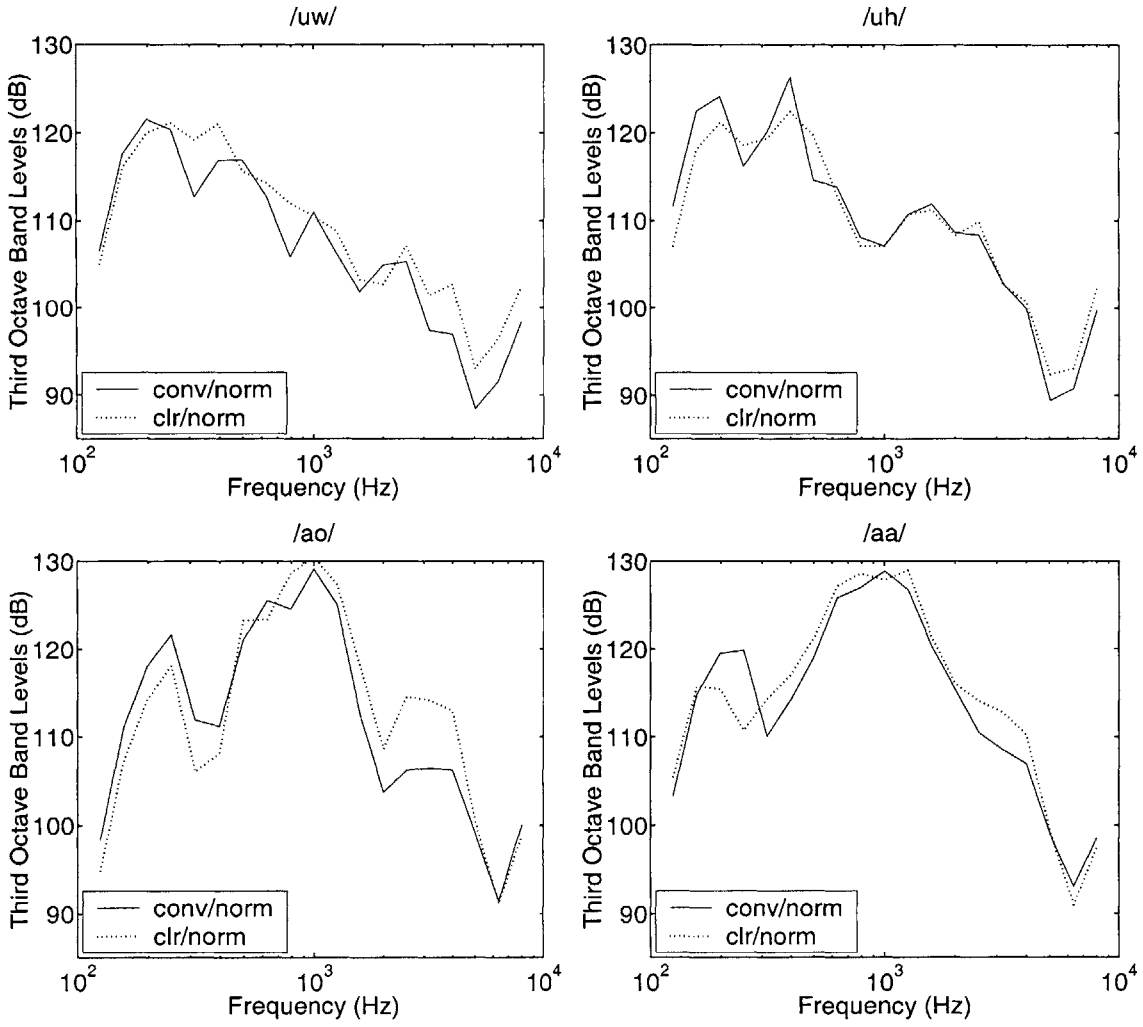

Figure A-5: Third-octave band spectra of low vowels in conv/normal and clear/normal modes for RG.

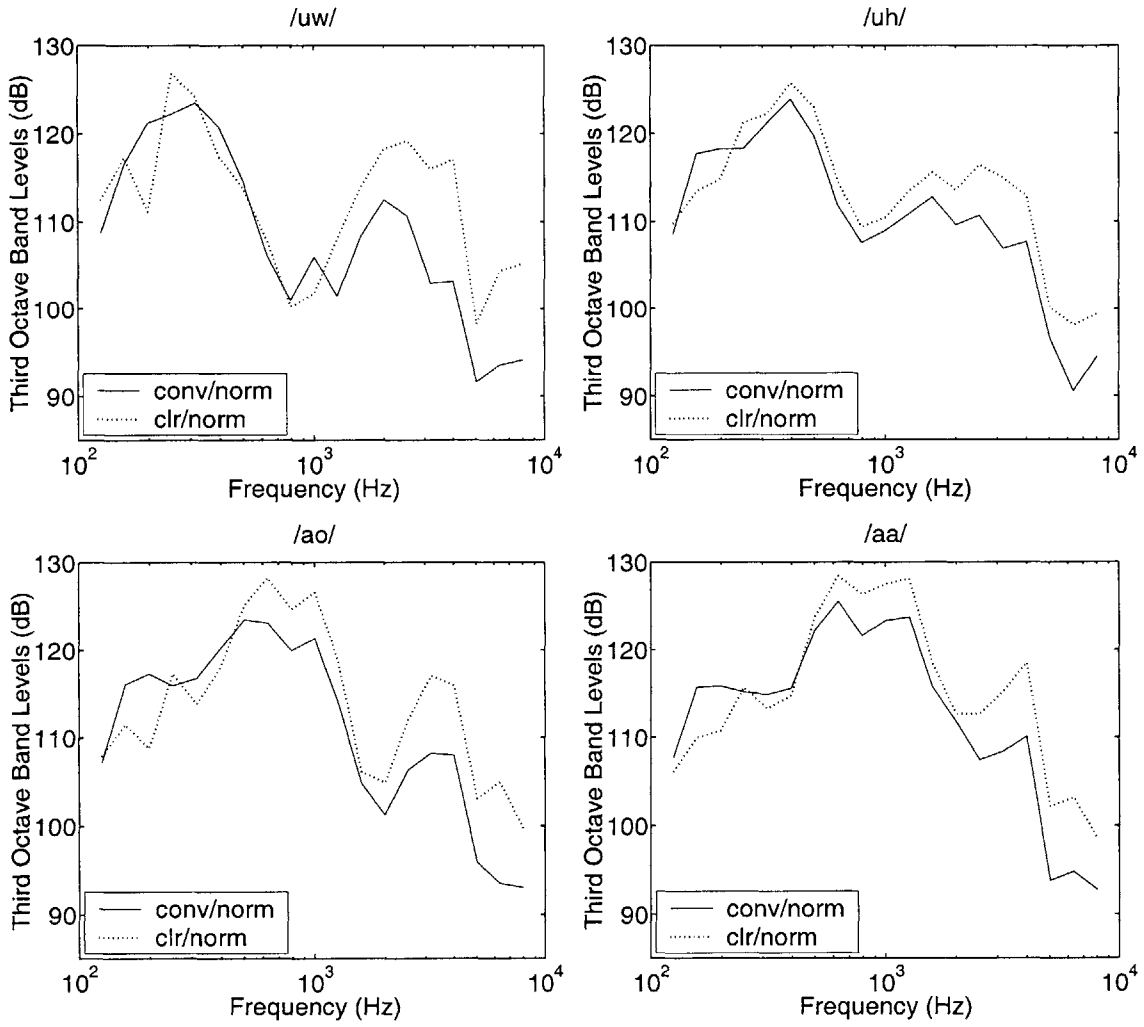

Figure A-6: Third-octave band spectra of low vowels in conv/normal and clear/normal modes for SA.

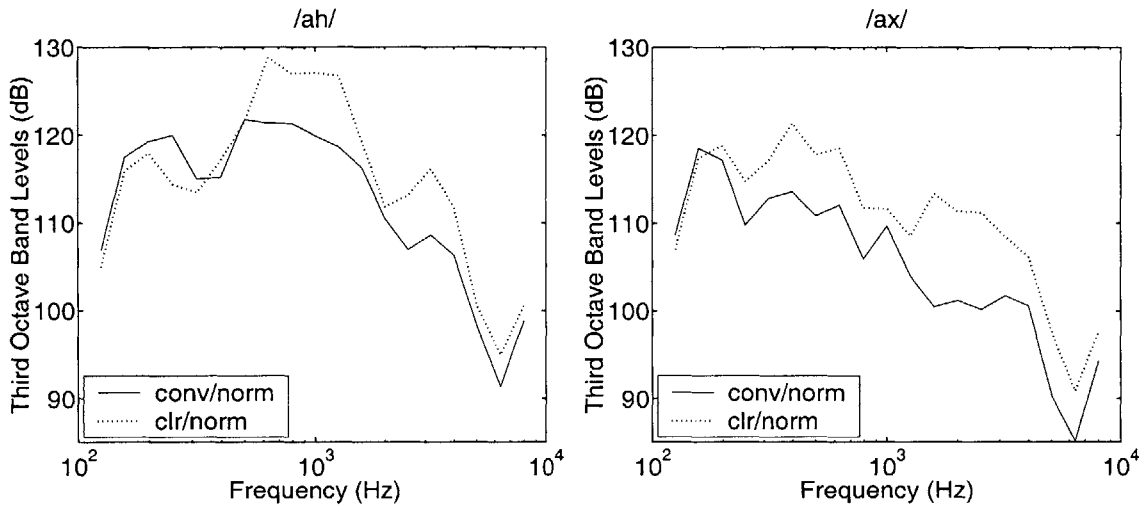

Figure A-7: Third-octave band spectra of neutral vowels in conv/normal and clear/normal modes for RG.

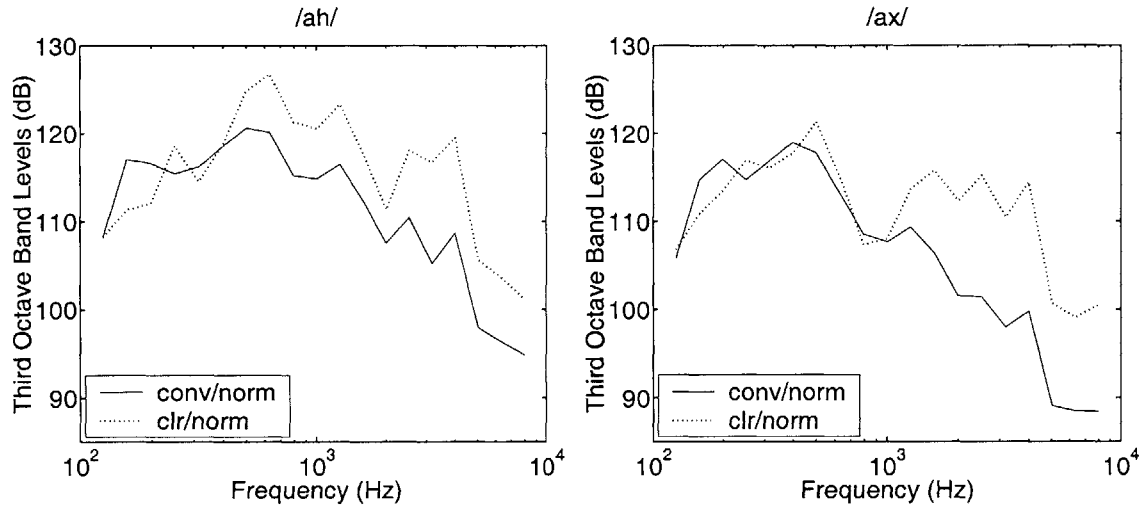

Figure A-8: Third-octave band spectra of neutral vowels in conv/normal and clear/normal modes for SA.

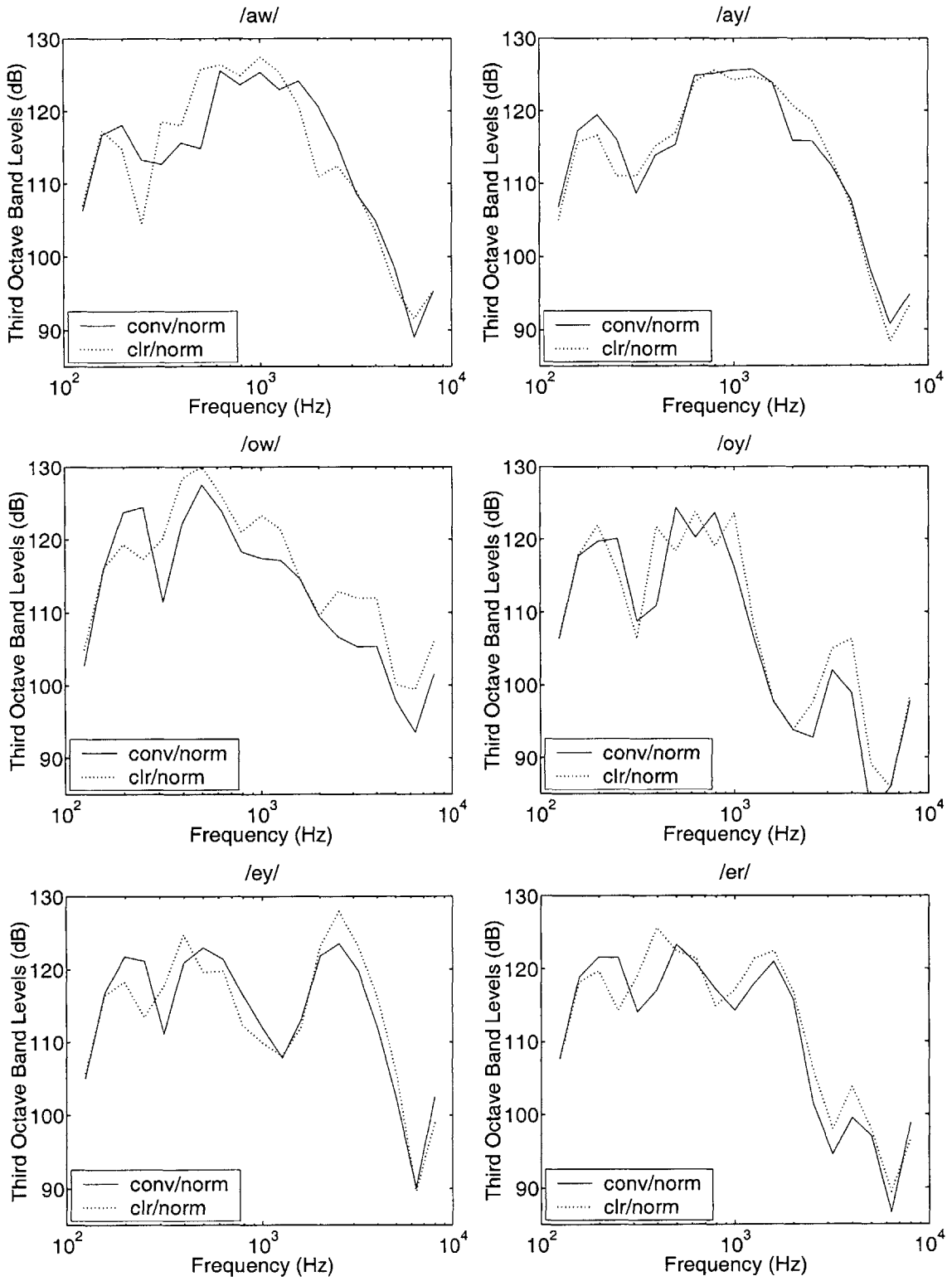

Figure A-9: Third-octave band spectra of diphthongs and retroflexed vowels in conv/normal and clear/normal modes for RG.

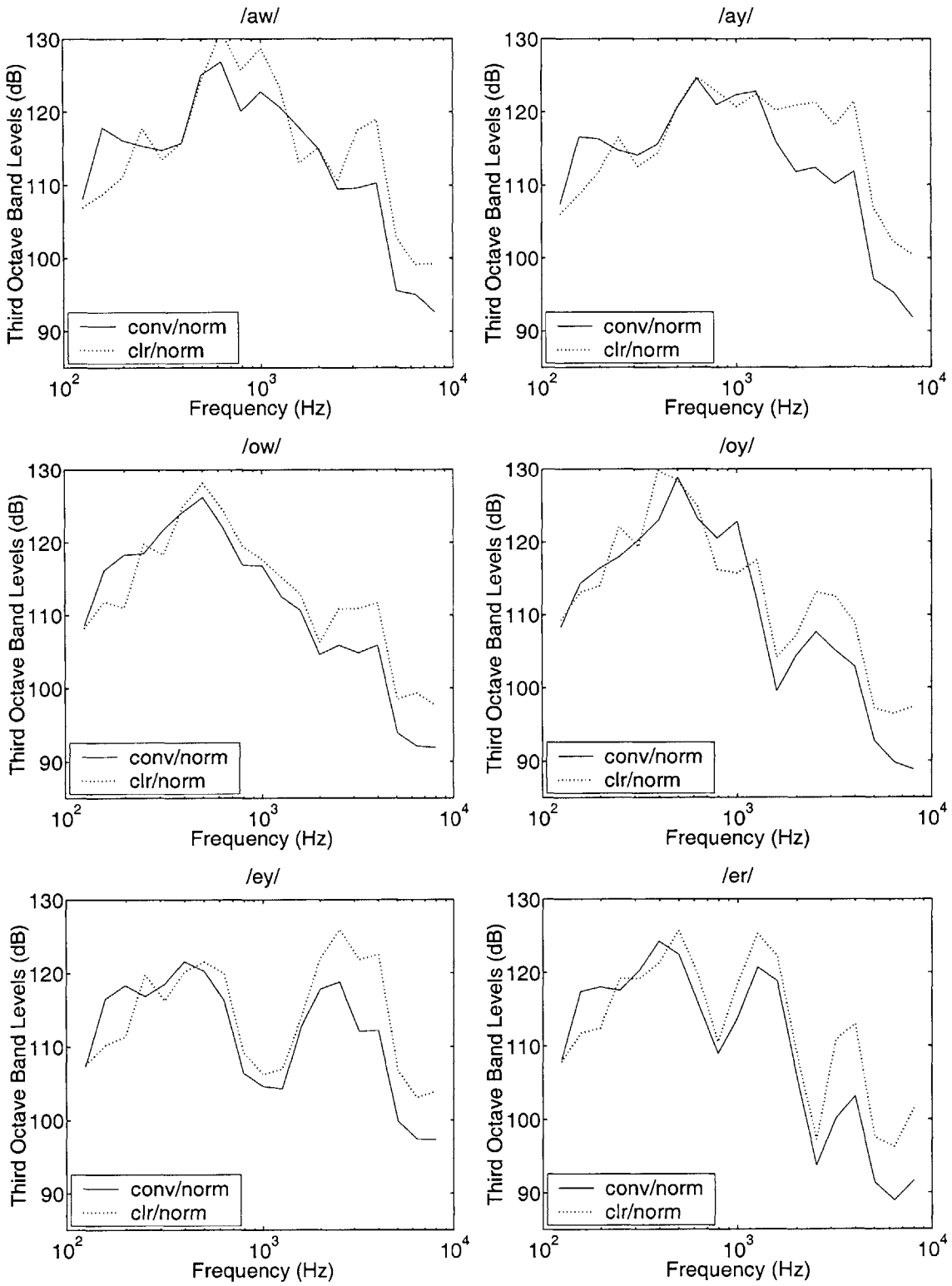

Figure A-10: Third-octave band spectra of diphthongs and retroflexed vowels in conv/normal and clear/normal modes for SA.

Appendix B

Acoustic Data for Processed Speech of RG and SA

B.1 Short-term Spectral Effects of Formant Processing

Figures B-1 through B-8 show the average short-term spectra of vowels from conv/normal, clear/normal, and formant processed/normal speech, normalized for segment RMS level for RG and SA. These figures show that the formant processing generally increased energy near second and third formant frequencies.

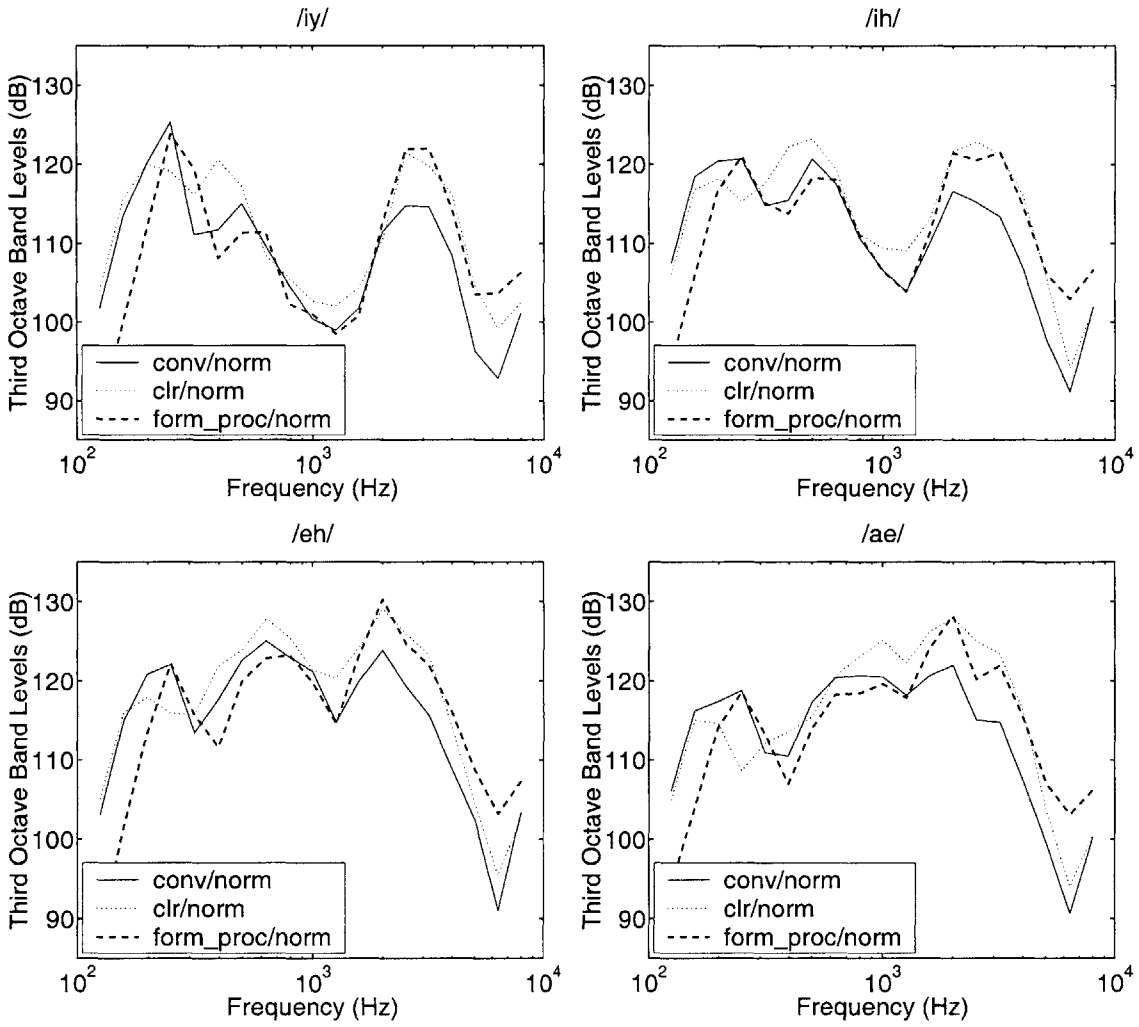

Figure B-1: Third-octave band spectra of high vowels in conv/normal, clear/normal, and (formant) processed/normal modes for RG.

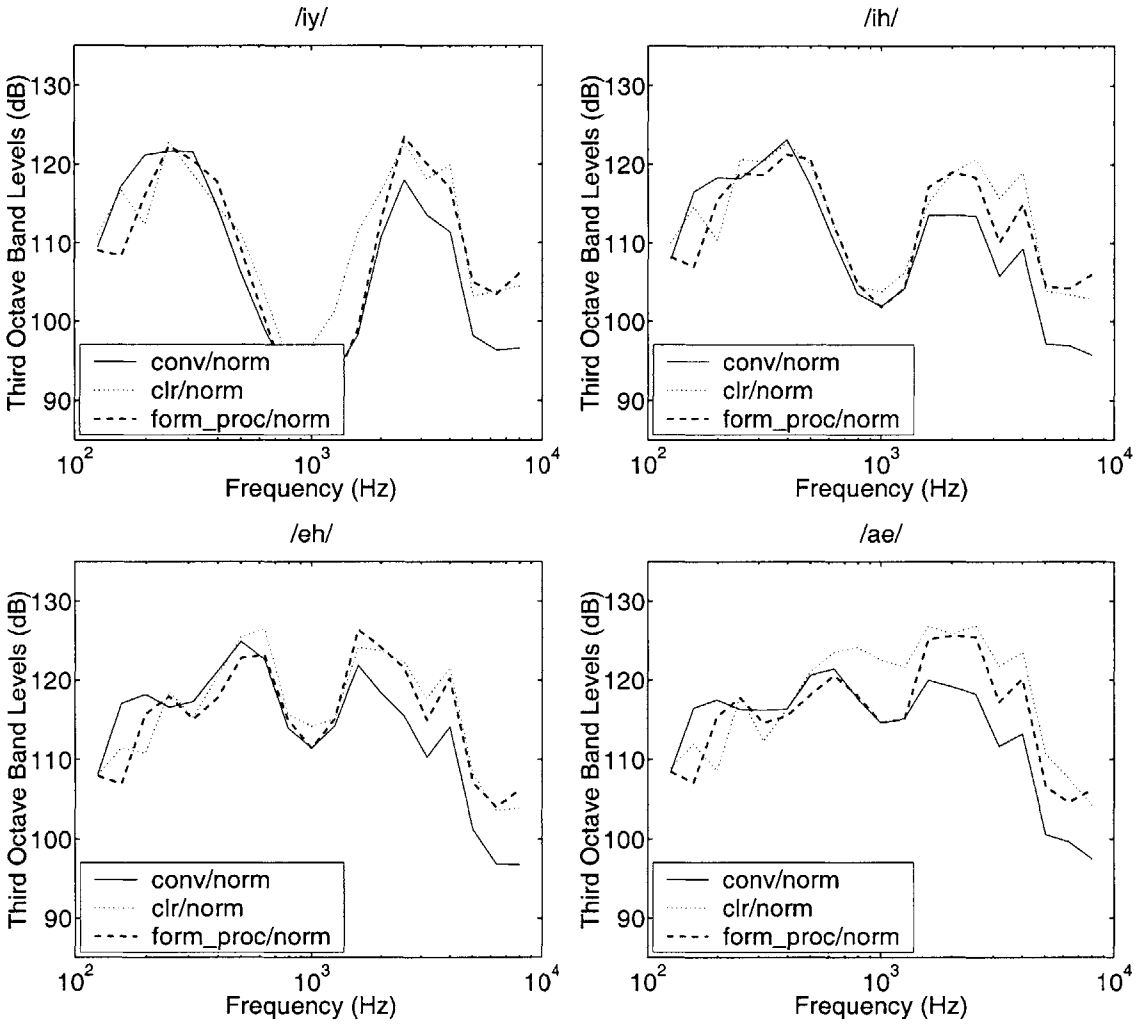

Figure B-2: Third-octave band spectra of high vowels in conv/normal, clear/normal and (formant) processed/normal modes for SA.

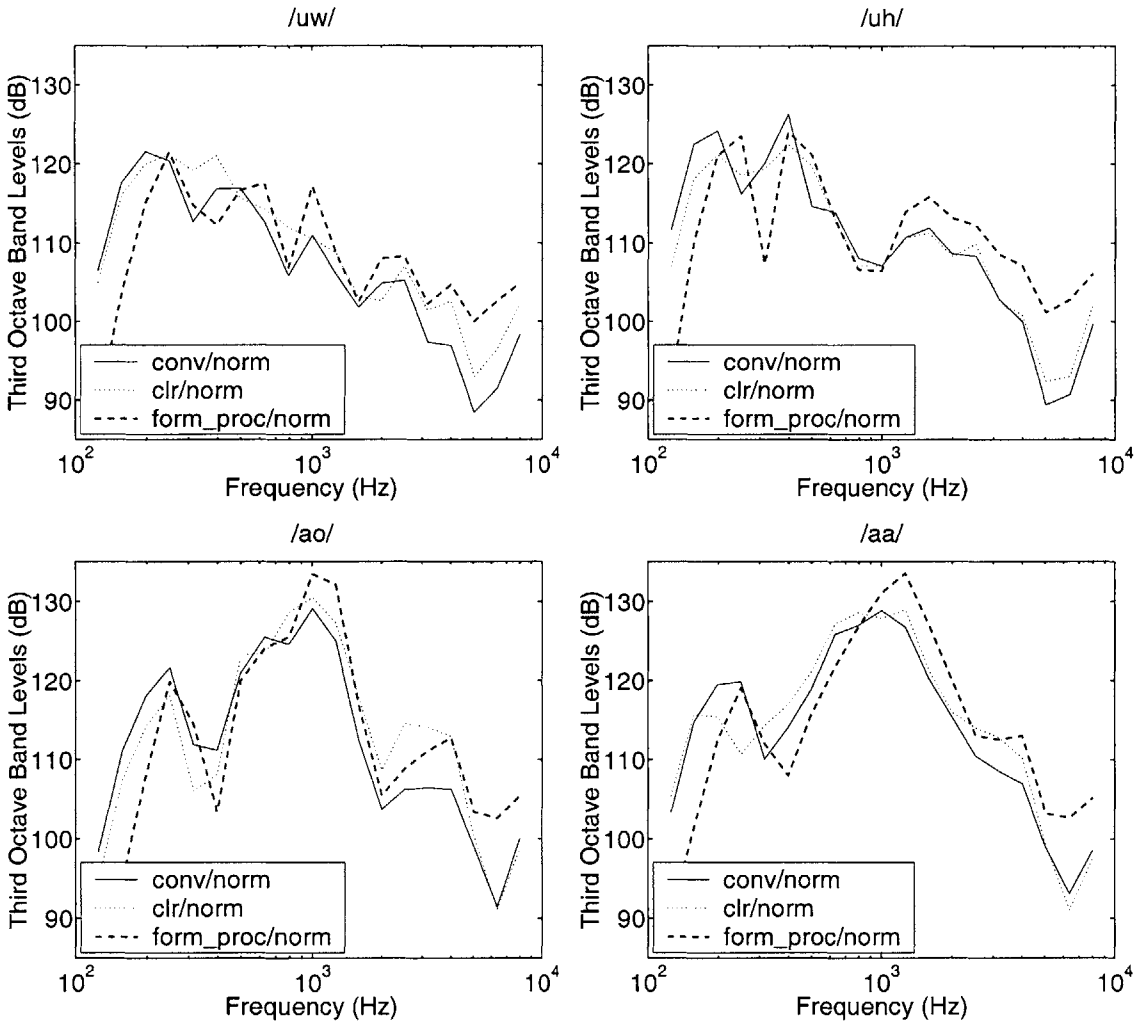

Figure B-3: Third-octave band spectra of low vowels in conv/normal, clear/normal and (formant) processed/normal modes for RG.

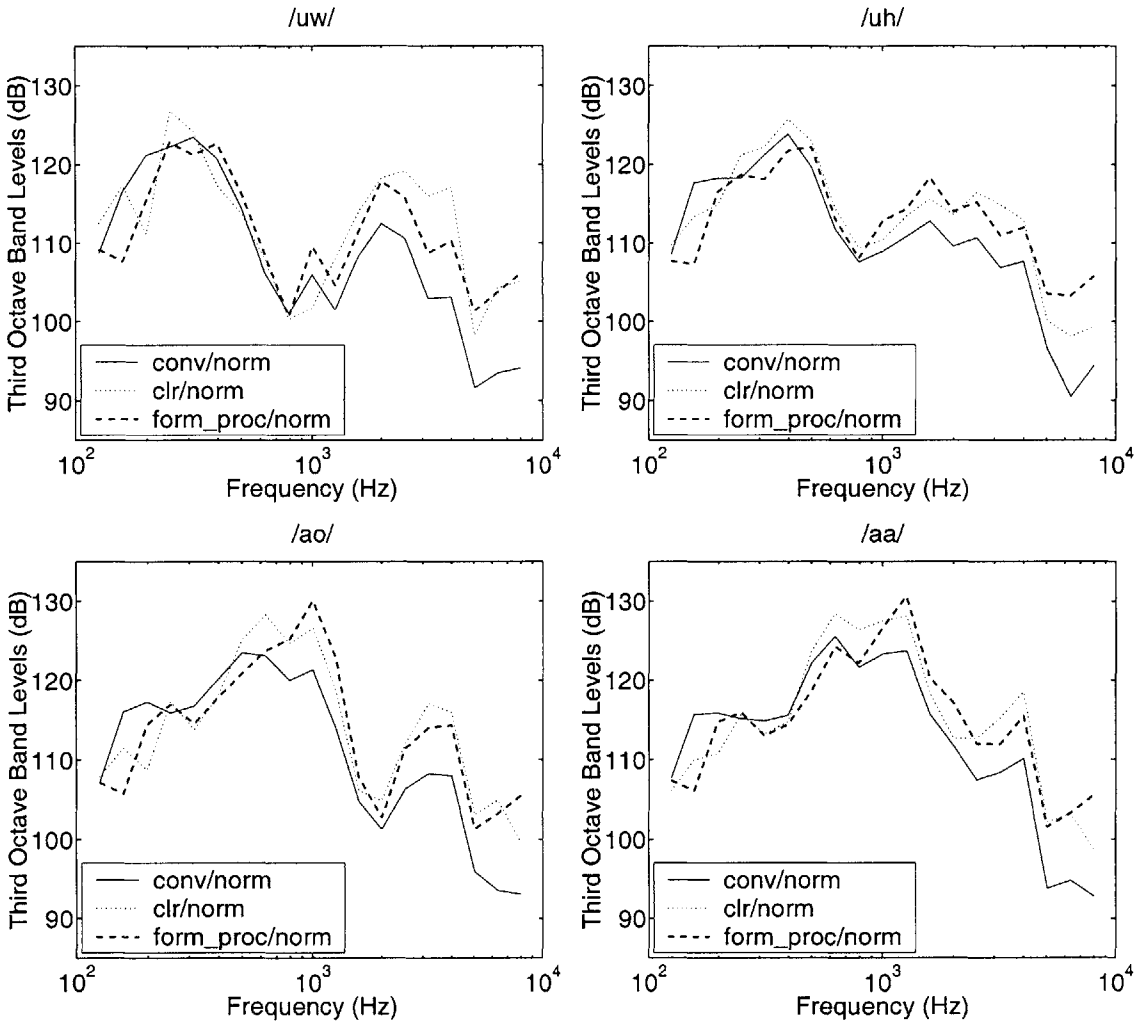

Figure B-4: Third-octave band spectra of low vowels in conv/normal, clear/normal and (formant) processed/normal modes for SA.

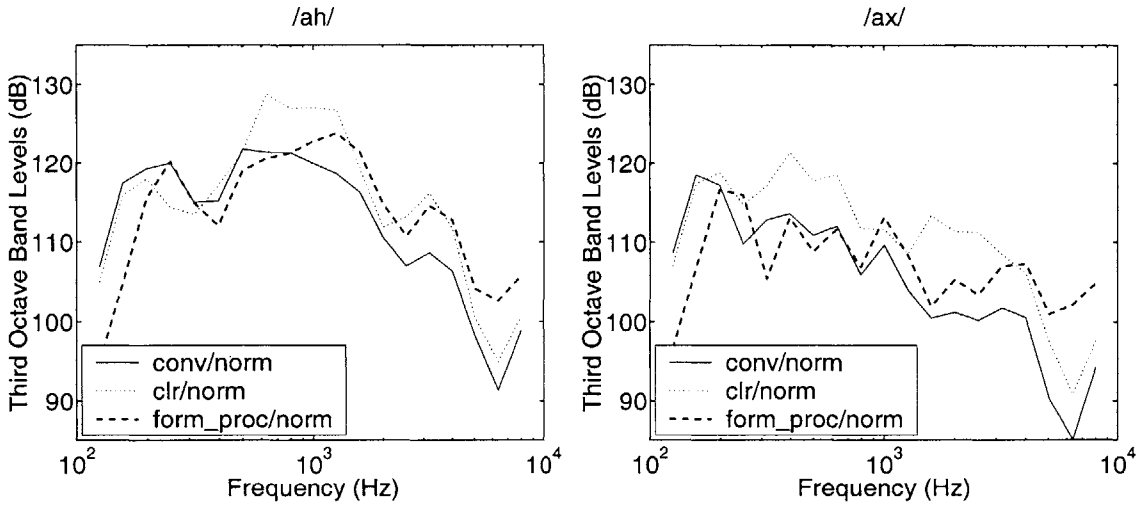

Figure B-5: Third-octave band spectra of neutral vowels in conv/normal, clear/normal and (formant) processed/normal modes for RG.

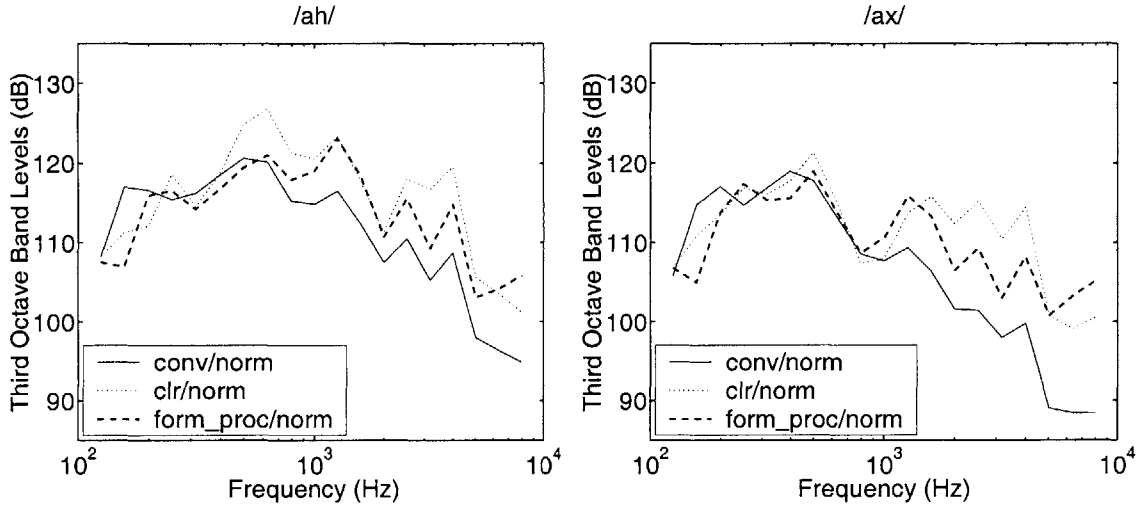

Figure B-6: Third-octave band spectra of neutral vowels in conv/normal, clear/normal and (formant) processed/normal modes for SA.

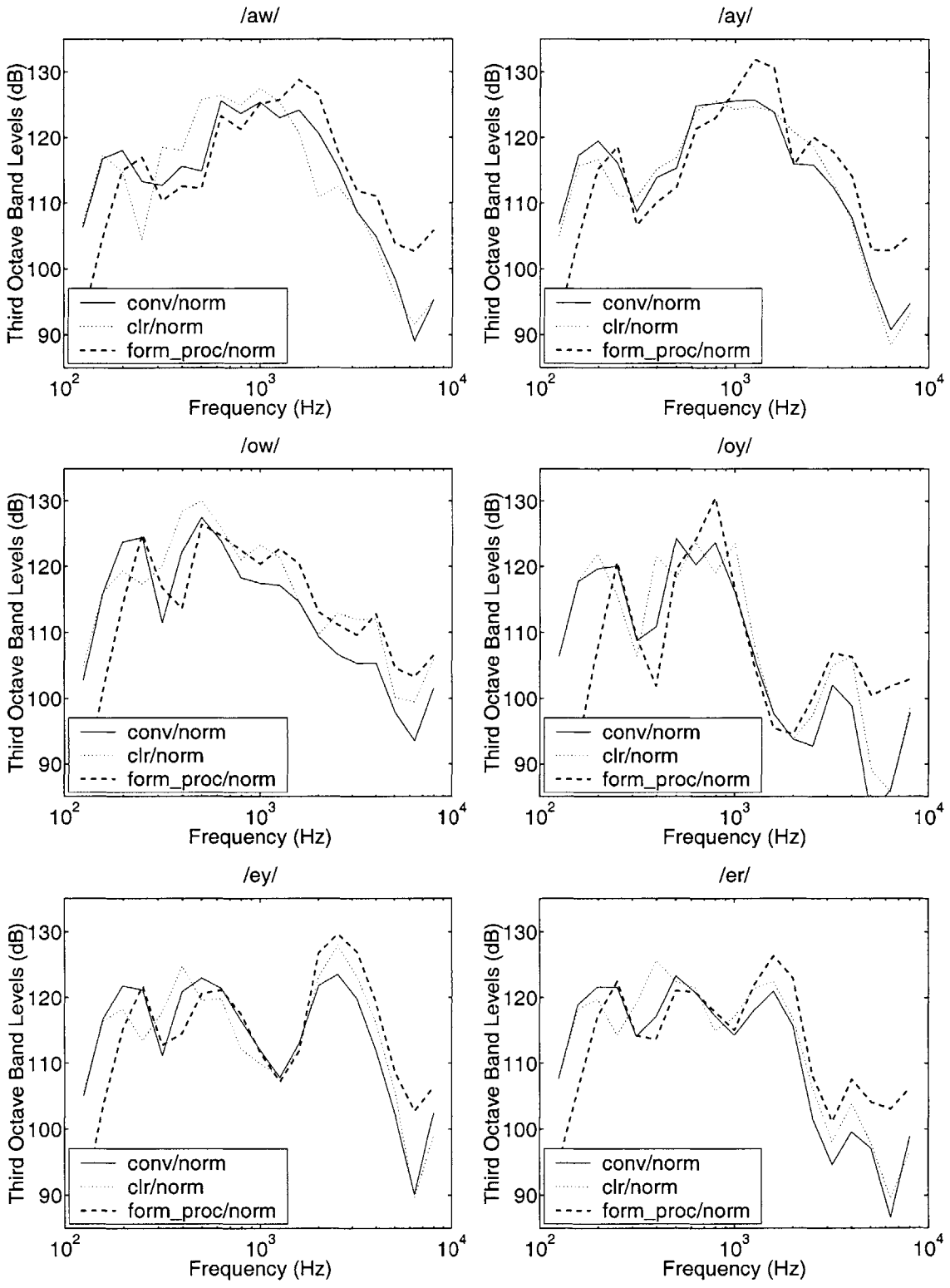

Figure B-7: Third-octave band spectra of diphthongs and retroflexed vowels in conv/normal, clear/normal, and (formant) processed/normal modes for RG.

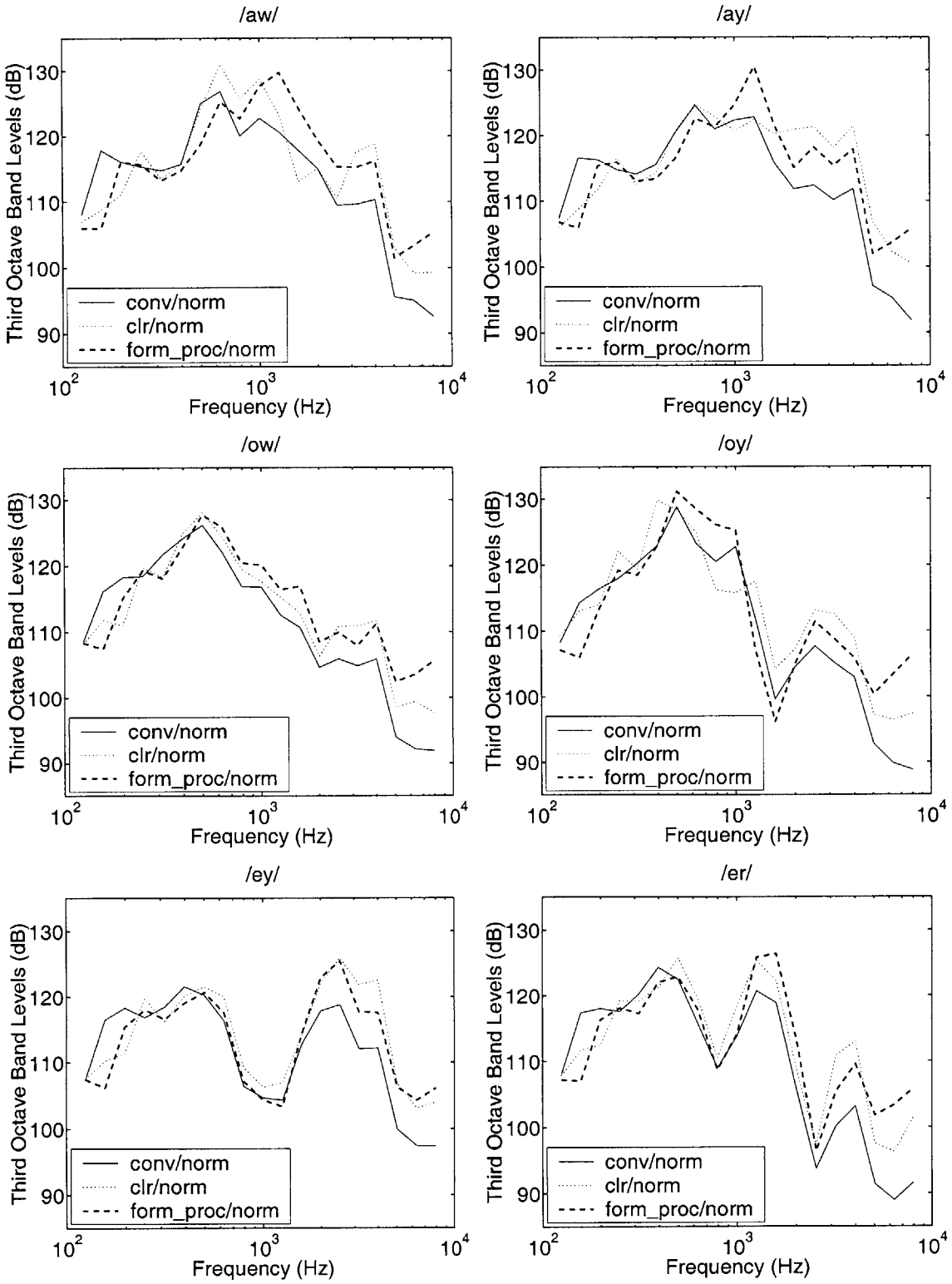

Figure B-8: Third-octave band spectra of diphthongs and retroflexed vowels in conv/normal, clear/normal, and (formant) processed/normal modes for SA.

Table B.1: Fundamental frequency distributions for the speech of SA and RG after applying the signal transformations in combination. Measurements were made, before and after processing, on the 30 sentences used as the combination stimuli for each talker in the first normal hearing intelligibility experiment.

Mode	RG		SA	
	Mean (Hz)	St Dev (Hz)	Mean (Hz)	St Dev (Hz)
conv/norm	210	48	102	15
combo_proc/norm	236	65	140	28

B.2 Fundamental Frequency Effects of Combination Processing

Table B.1 shows the average and standard deviation of fundamental frequency for RG and SA, before and after applying the signal transformations in combination.

B.3 Long-term Spectral Effects of Other Transformations

Figures B-9 and B-10 show the long-term spectra of envelope processed speech relative to the long term spectra of conversational speech for RG and SA. Envelope processing had no substantial effect on the long term spectra.

Pitch processing, however, did affect the long-term spectra to some extent. This is explained in detail in Chapter 5.

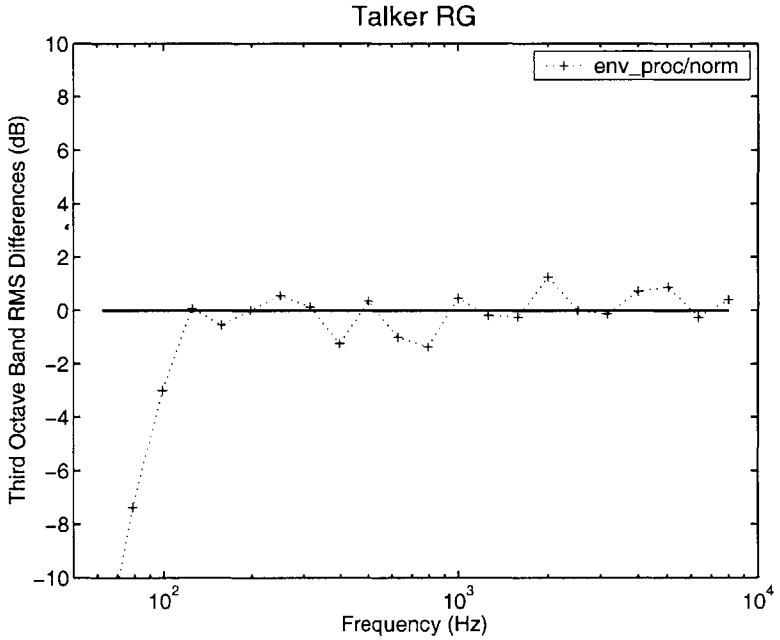

Figure B-9: Third-octave band RMS spectral differences of RG's clear/normal and (envelope) processed/normal modes relative to conv/normal speech. The processing did not affect the long-term spectrum significantly.

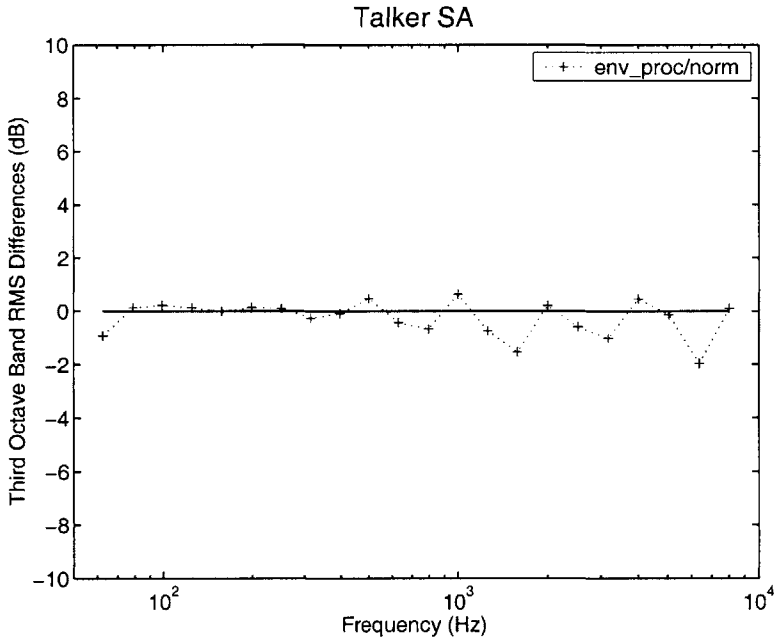

Figure B-10: Third-octave band RMS spectral differences of SA's clear/normal and (envelope) processed/normal modes relative to conv/normal speech. The processing did not affect the long-term spectrum significantly.

Table B.2: Fundamental frequency distributions for SA and RG. For a given talker, measurements for each condition were made on the same set of 50 sentences.

Mode	RG		SA	
	Mean (Hz)	St Dev (Hz)	Mean (Hz)	St Dev (Hz)
conv/norm	216	49	105	18
clear/slow	196	49	154	36
clear/norm	209	40	138	25
pitch_proc/norm	247	67	150	31
form_proc/norm	215	49	106	19
env_proc/norm	207	55	106	17

B.4 Fundamental Frequency Effects of Other Transformations

Table B.2 shows the average and standard deviation of fundamental frequency for RG and SA in each condition. This Table shows that neither formant nor envelope processing altered F0 average or range.

B.5 Temporal Envelope Effects of Other Transformations

Figures B-11 and B-12 show the intensity envelope spectra of formant processed speech for RG and SA. Formant processing had no substantial effect on the intensity envelope spectra of speech.

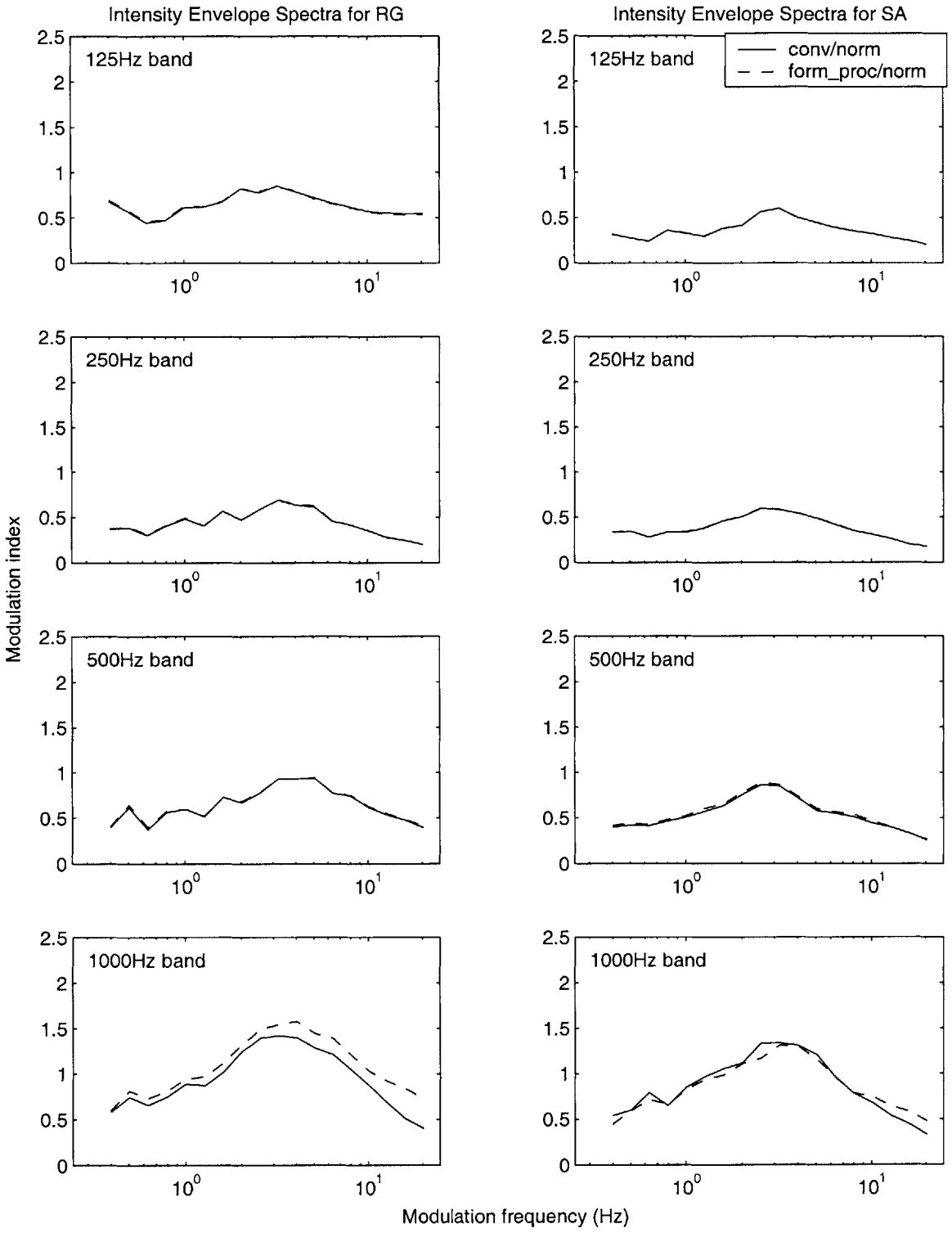

Figure B-11: Spectra of intensity envelopes, before and after applying the formant processing, for Talkers RG and SA in lower four octave bands.

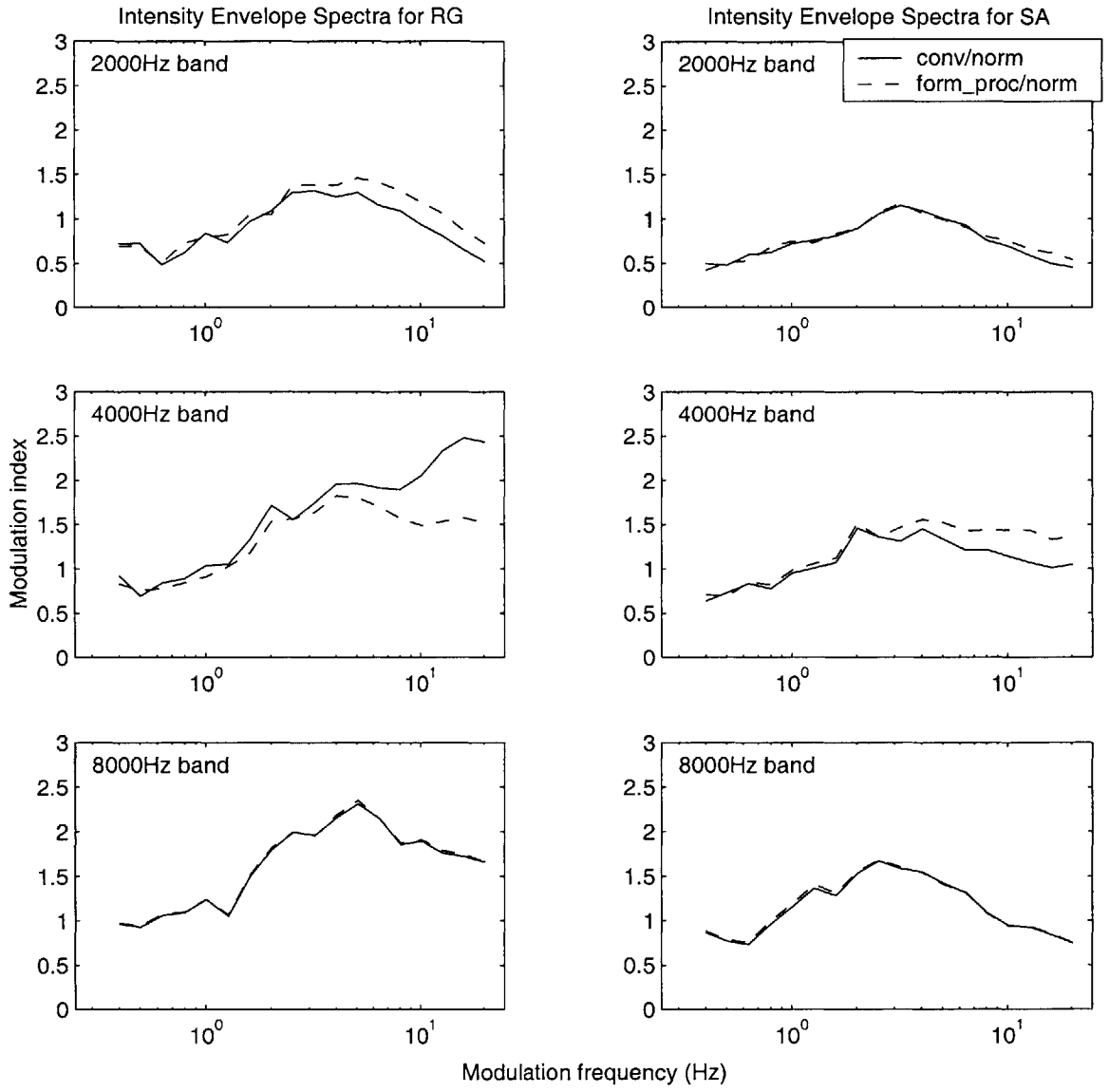

Figure B-12: Spectra of intensity envelopes, before and after applying formant processing, for Talkers RG and SA in upper three octave bands.

Table B.3: Fundamental frequency distributions for RG and SA, before and after LPC processing without altering the pitch.

Mode	RG		SA	
	Mean (Hz)	St Dev (Hz)	Mean (Hz)	St Dev (Hz)
conv/norm	216	50	106	14
proc_same_pitch/norm	219	52	107	17

B.6 Processing Used to Assess Signal Processing Artifacts

Values of F0 average and standard deviation are provided in Table B.3 after LPC analysis-synthesis processing, without F0 modification. F0 distribution of unprocessed speech is provided for reference. The processed(same_pitch) speech had the same F0 contour as conversational speech.

Figures B-13 and B-14 show the long-term spectra of RG and SA for both the twice-processed(same_formant) and the processed(enhanced_formant) conditions. From these graphs, it appears that the twice-processing scheme was successful in restoring formants to their original values.

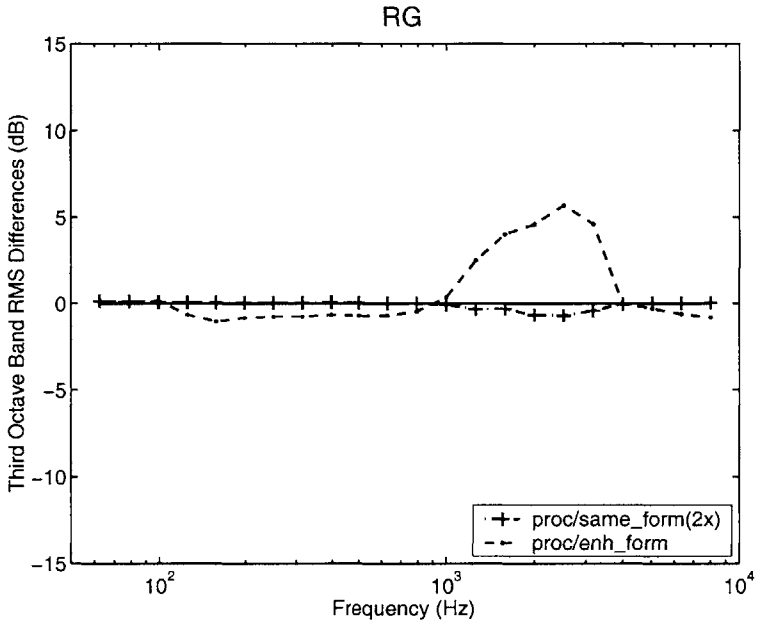

Figure B-13: Third-octave band RMS spectral differences of RG's processed(same_formant)/normal and processed(enhanced_formant)/normal modes relative to conv/normal speech.

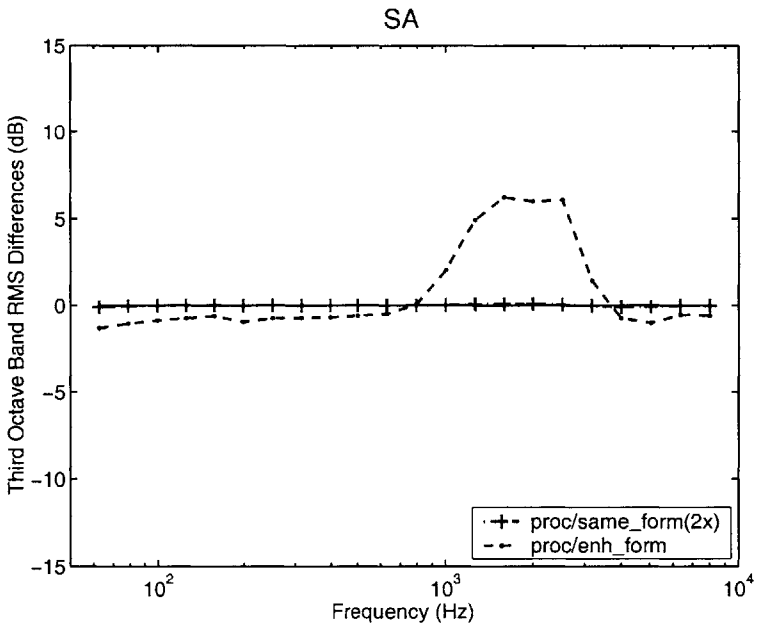

Figure B-14: Third-octave band RMS spectral differences of SA's processed(same_formant)/normal and processed(enhanced_formant)/normal modes relative to conv/normal speech.

Figures B-15 and B-16 show the intensity envelope spectra of RG and SA for both the twice-processed(same_envelope) and the processed(enhanced_envelope) conditions. From these graphs, it appears that the twice-processing scheme was successful in restoring envelopes to their original values.

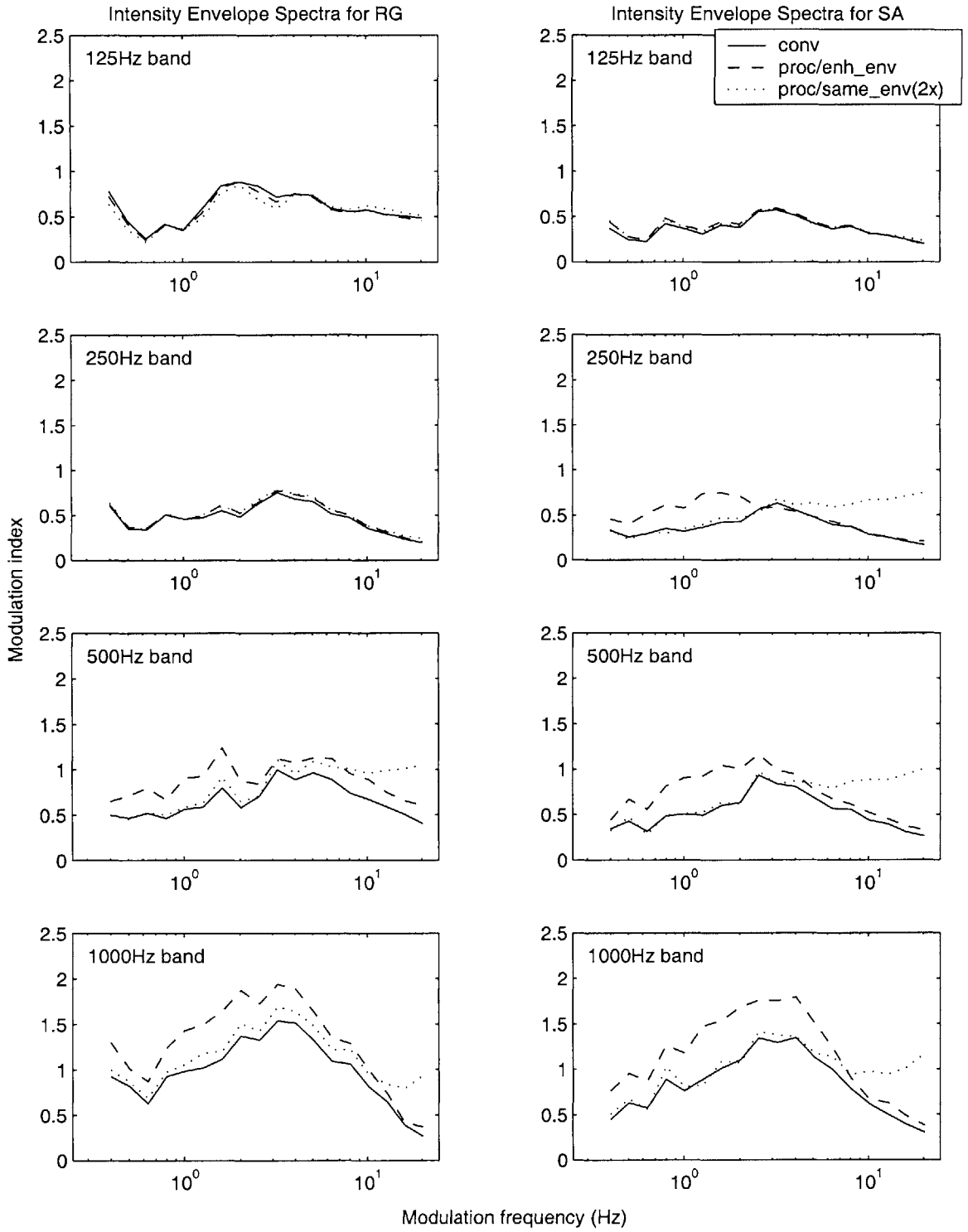

Figure B-15: Spectra of intensity envelopes for Talkers RG and SA in lower four octave bands for processed(same_envelope)/normal and processed(enhanced_envelope)/normal modes. Spectra for conversational envelopes is provided as a reference.

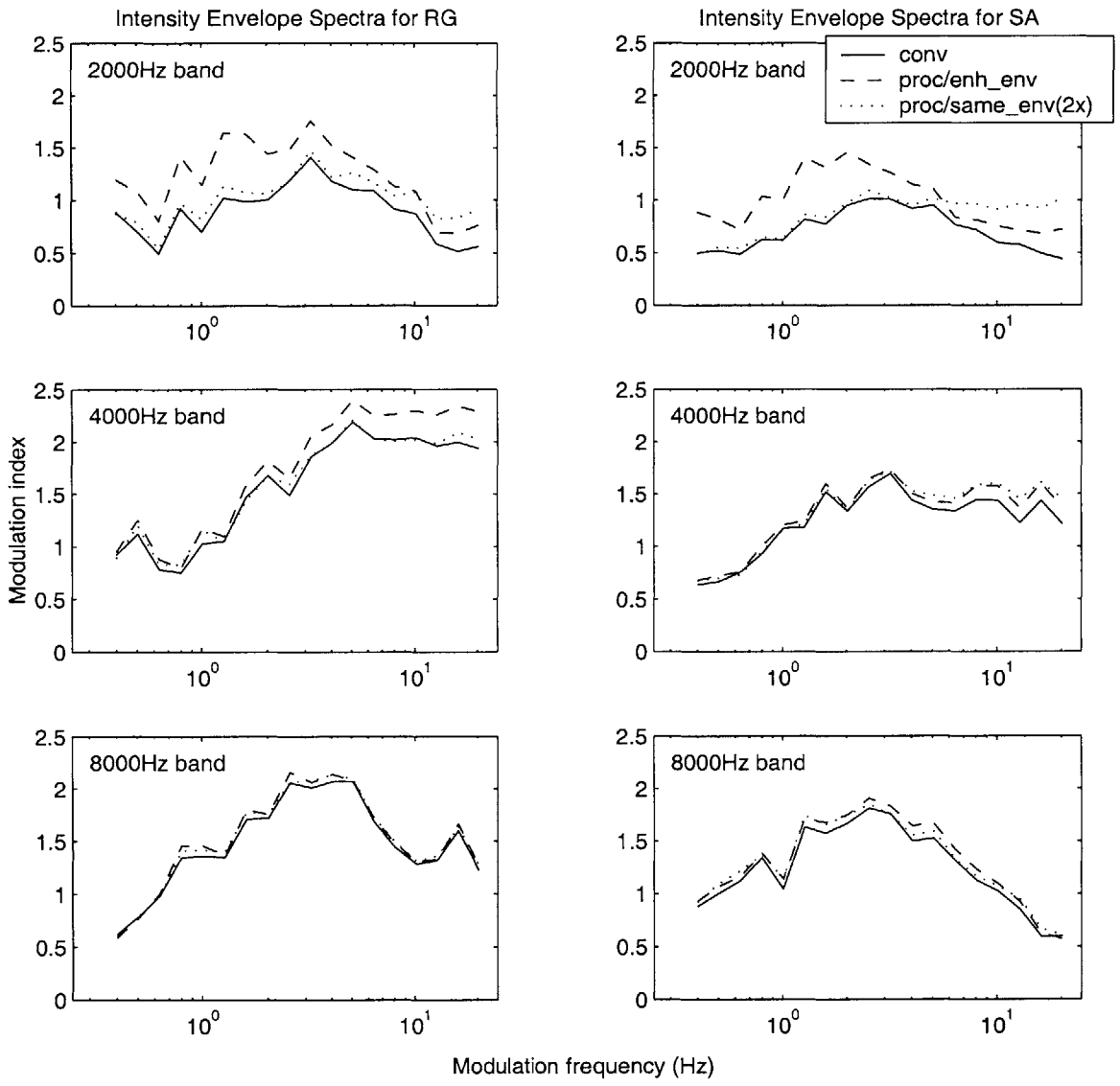

Figure B-16: Spectra of intensity envelopes for Talkers RG and SA in upper three octave bands for processed(same_envelope)/normal and processed(enhanced_envelope)/normal modes. Spectra for conversational envelopes is provided as a reference.

Appendix C

Acoustic Data for Processed Speech of T3 and T4

C.1 Processing of Formant Frequencies

Figures C-1 and C-2 show the effect of formant processing on the long-term spectra of speech from T3 and T4. As with RG and SA, the processing had the desired effect of increasing energy above 1kHz.

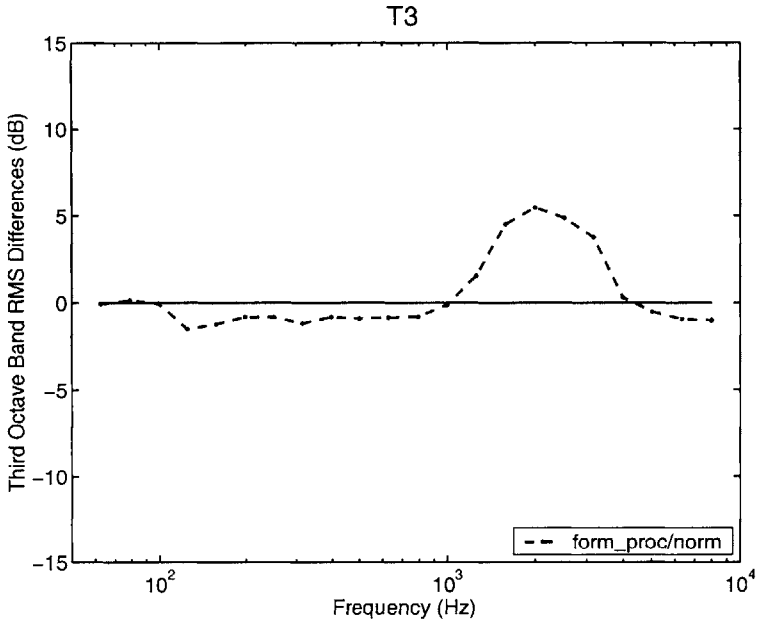

Figure C-1: Third-octave band RMS spectral differences of T3's clear/normal and (formant) processed/normal modes relative to conv/normal speech.

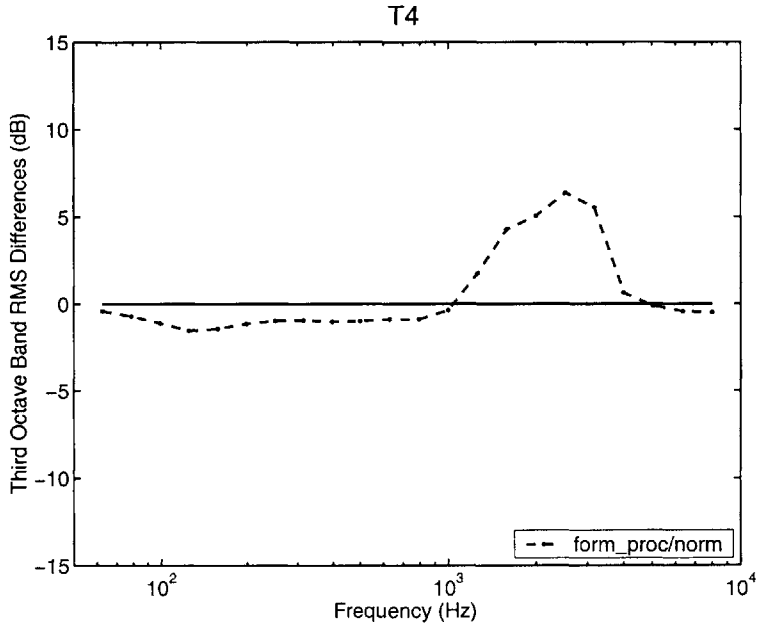

Figure C-2: Third-octave band RMS spectral differences of T4's clear/normal and (formant) processed/normal modes relative to conv/normal speech.

C.2 Processing of Fundamental Frequency

Figure C-3 shows the effect of pitch processing on the fundamental frequency distribution of speech from T3 and T4. As with RG and SA, the processing had the desired effect of increasing F0 average and range.

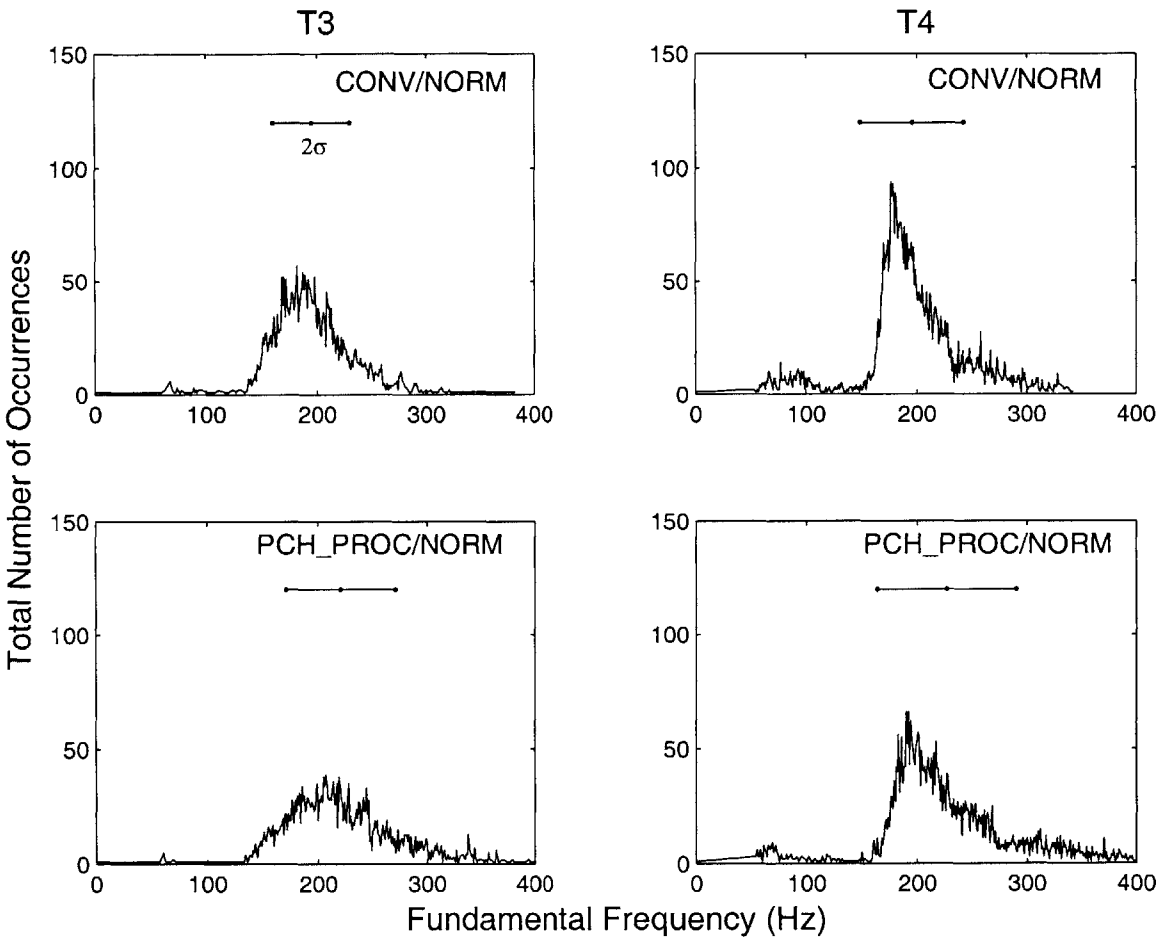

Figure C-3: Fundamental frequency distributions for conv/normal and (pitch) processed/normal speech of T3 and T4. Each row shows distributions for different speaking modes; columns give results for each talker.

C.3 Processing of Temporal Envelopes

Figures C-4 and C-5 show the effect of envelope processing on the intensity envelope spectra of speech from T3 and T4. As with RG and SA, the processing had the desired effect of increasing low frequency modulations.

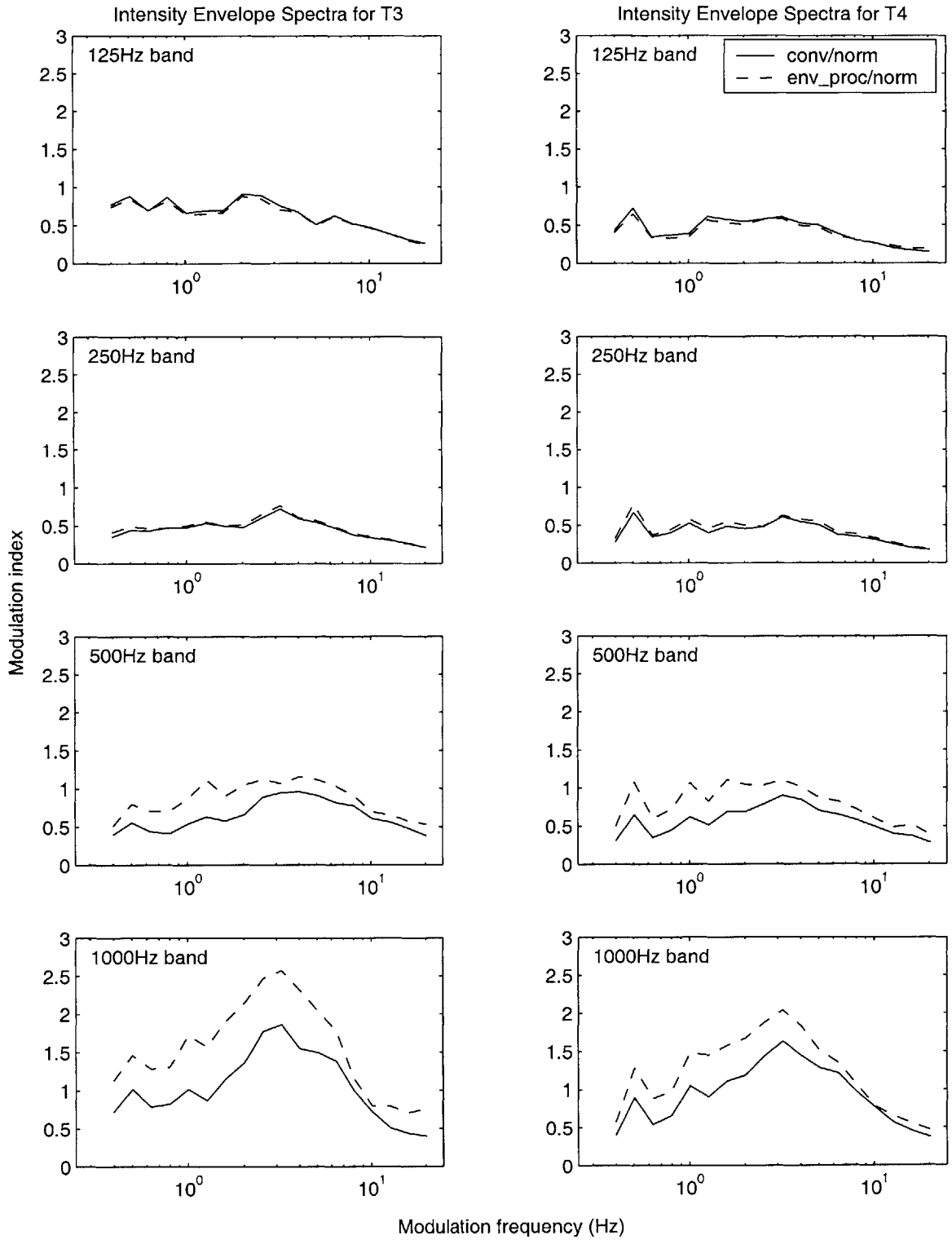

Figure C-4: Spectra of intensity envelopes for Talkers T3 and T4 in lower four octave bands.

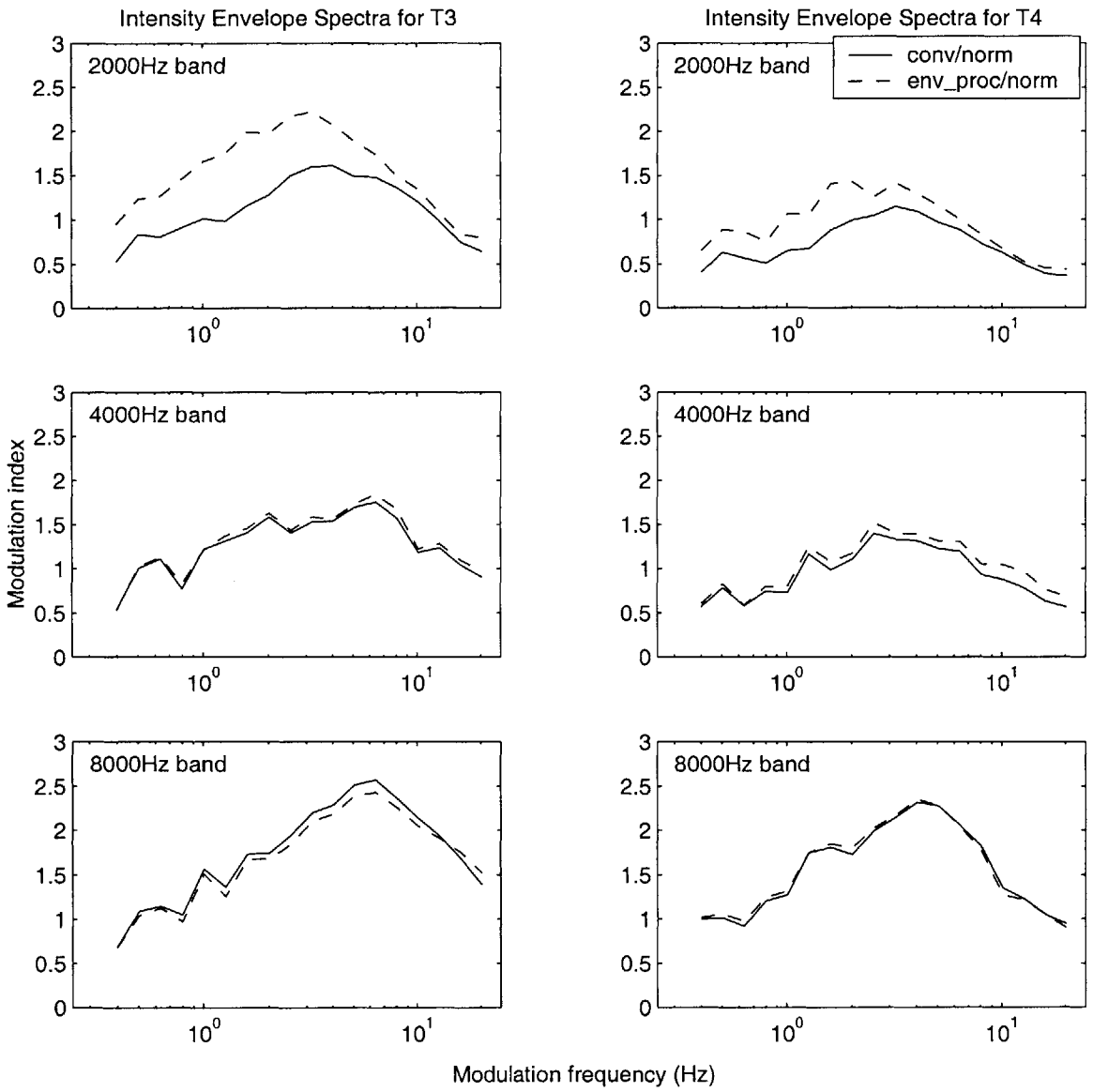

Figure C-5: Spectra of intensity envelopes for Talkers T3 and T4 in upper three octave bands.

C.4 Effect of Other Transformations on Long-term Spectra

Figures C-6 and C-7 show the long-term spectra of envelope processed speech relative to the long term spectra of conversational speech for T3 and T4. As with RG and SA, envelope processing had no substantial effect on the long term spectra.

Figures C-8 and C-9 show the long-term spectra of pitch processed speech relative to the long term spectra of conversational speech for T3 and T4. As with RG and SA, pitch processing did produce a high-frequency emphasis of roughly 2 to 4 dB.

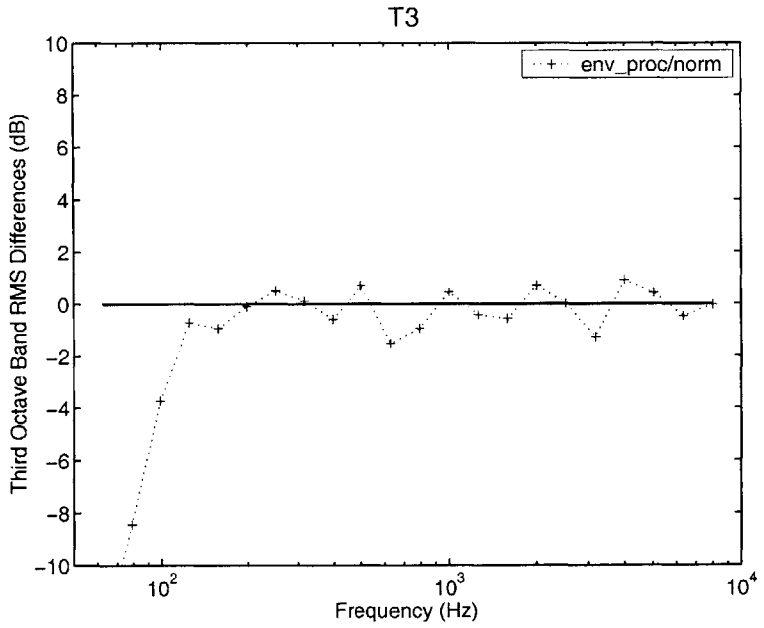

Figure C-6: Third-octave band RMS spectral differences of T3's clear/normal and (envelope) processed/normal modes relative to conv/normal speech. The processing had no substantial effect on the long-term spectrum.

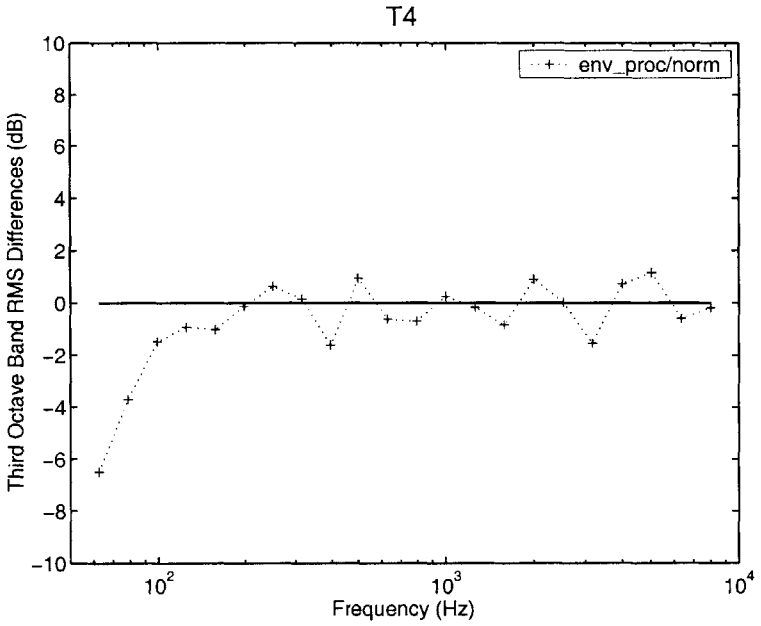

Figure C-7: Third-octave band RMS spectral differences of T4's clear/normal and (envelope) processed/normal modes relative to conv/normal speech. The processing had no substantial effect on the long-term spectrum.

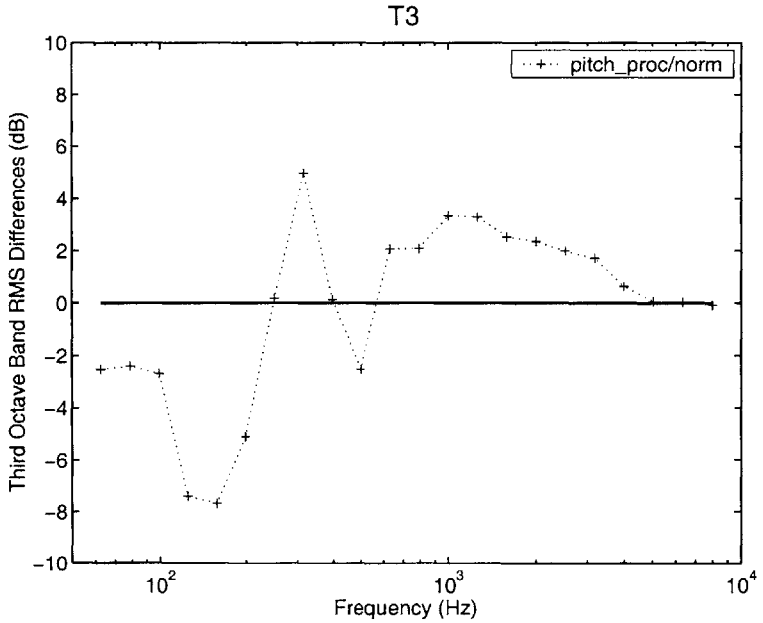

Figure C-8: Third-octave band RMS spectral differences of T3's clear/normal and (pitch) processed/normal modes relative to conv/normal speech. The processing resulted in a small high-frequency emphasis above 1kHz.

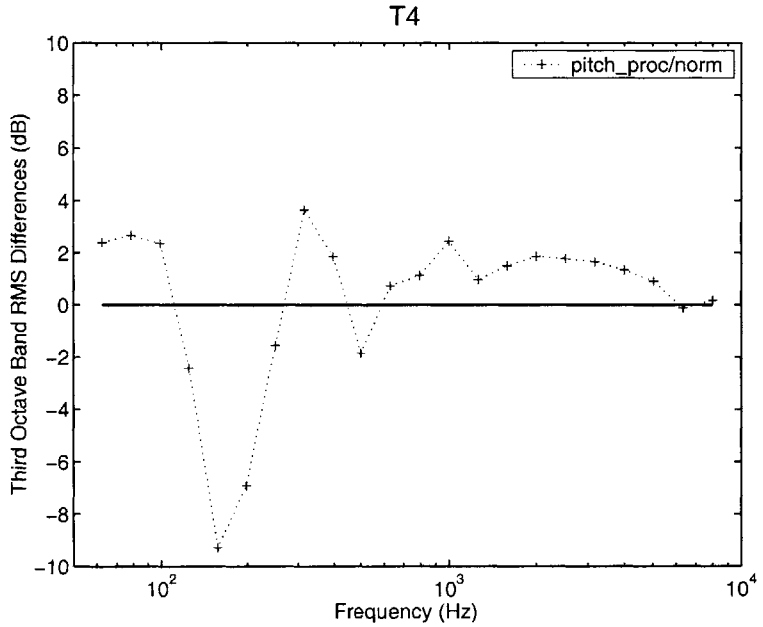

Figure C-9: Third-octave band RMS spectral differences of T4's clear/normal and (pitch) processed/normal modes relative to conv/normal speech. The processing resulted in a small high-frequency emphasis above 1kHz.

Table C.1: Fundamental frequency distributions for T3 and T4. For a given talker, measurements for each condition were made on set of 30 sentences used as stimuli for that condition in the first normal hearing intelligibility experiment. Measurements prior to processing are provided for reference.

Mode	T3		T4	
	Mean (Hz)	St Dev (Hz)	Mean (Hz)	St Dev (Hz)
conv/norm	185	29	198	50
form_proc/norm	185	29	198	49
conv/norm	195	35	196	47
pitch_proc/norm	221	50	227	63
conv/norm	204	33	193	51
env_proc/norm	197	41	191	52

C.5 Effect of Other Transformations on Fundamental Frequency

Table C.1 shows the average and standard deviation of fundamental frequency for T3 and T4 in each condition. This Table shows that neither formant nor envelope processing altered F0 average or range.

C.6 Effect of Other Transformations on Temporal Envelopes

Figures C-10 and C-11 show the intensity envelope spectra of formant processed speech for T3 and T4. As with RG and SA, formant processing had no substantial effect on the intensity envelope spectra of speech.

Figures C-12 and C-13 show the intensity envelope spectra of pitch processed speech for T3 and T4. Although to a lesser extent than with RG and SA, pitch processing did increase the modulation index of high frequency modulations for T3 in the 250Hz band and for T4 in the 125Hz, 250Hz, and 1000Hz bands.

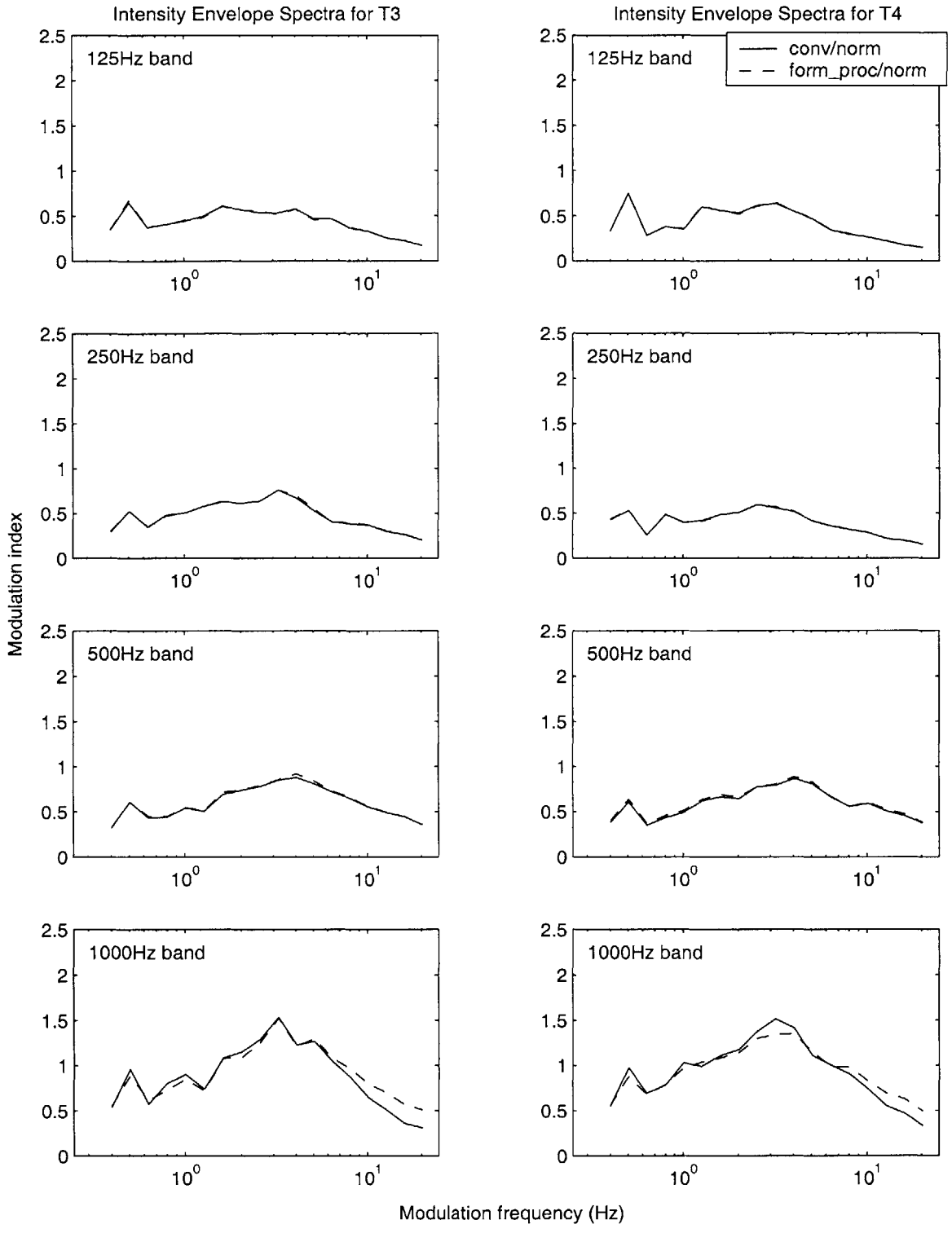

Figure C-10: Spectra of intensity envelopes for Talkers T3 and T4 in lower four octave bands, before and after (formant) processing.

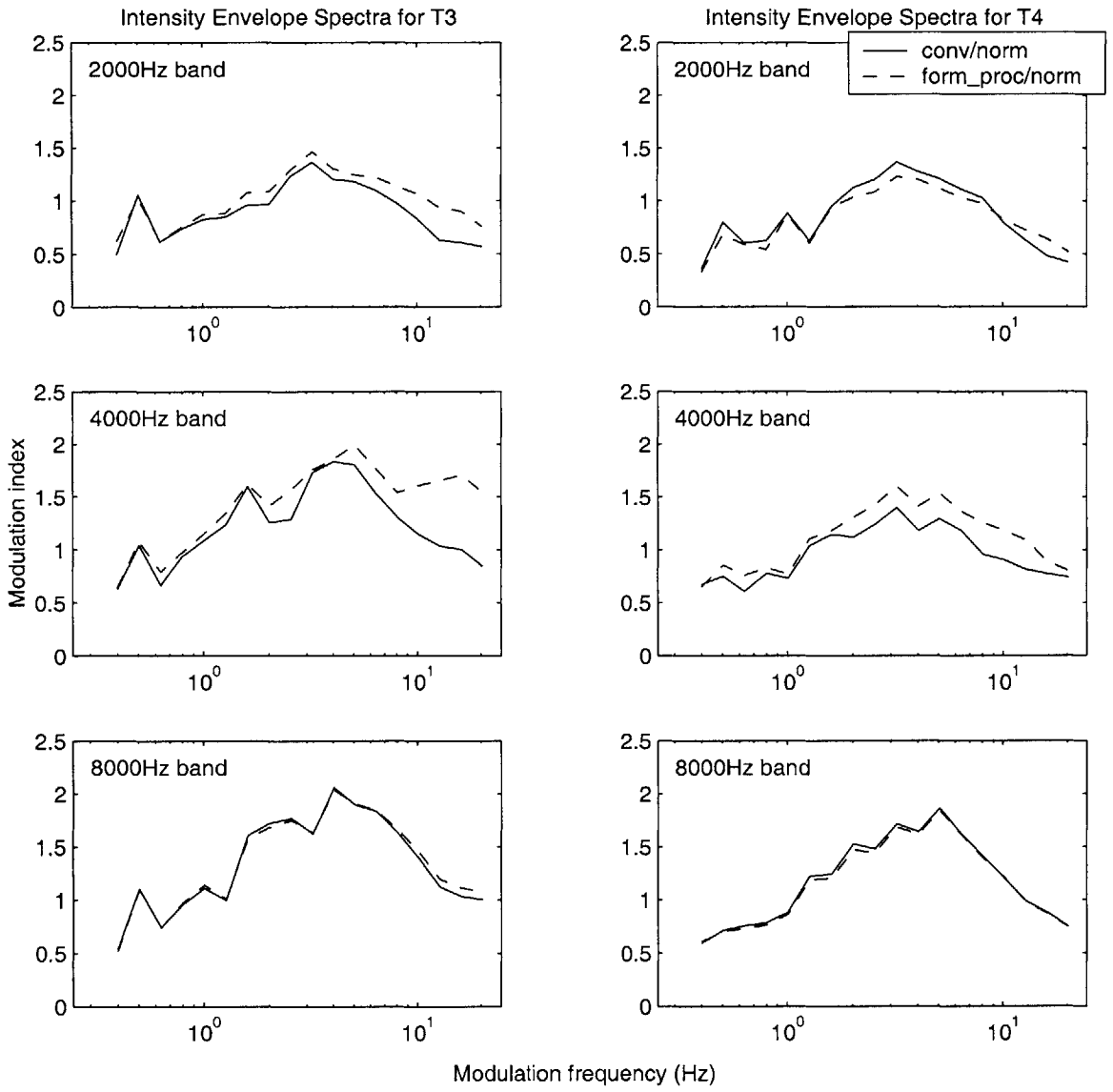

Figure C-11: Spectra of intensity envelopes for Talkers T3 and T4 in upper three octave bands, before and after (formant) processing.

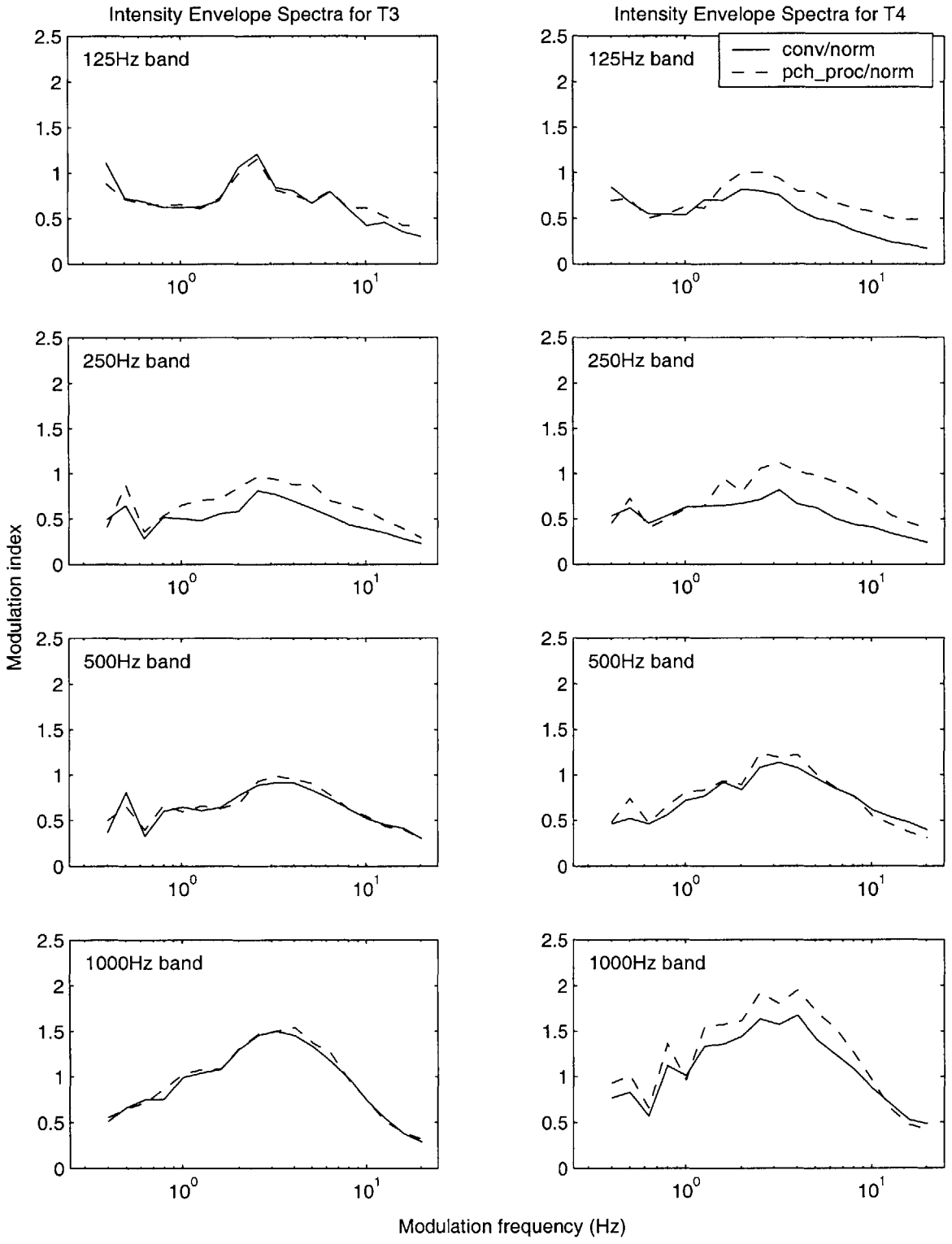

Figure C-12: Spectra of intensity envelopes for Talkers T3 and T4 in lower four octave bands, before and after (pitch) processing.

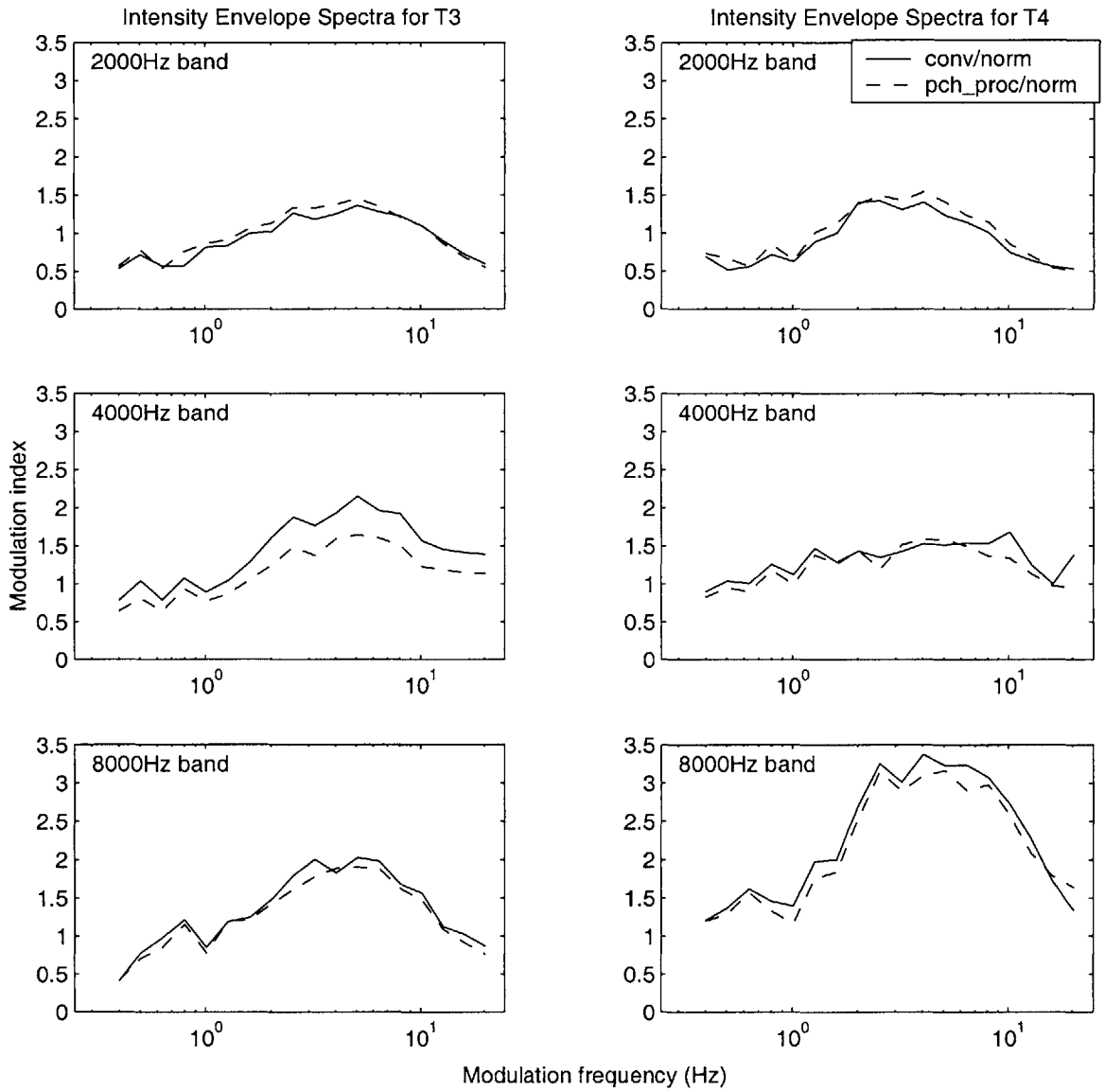

Figure C-13: Spectra of intensity envelopes for Talkers T3 and T4 in upper three octave bands, before and after (pitch) processing.

C.7 Processing Used to Assess Signal Processing Artifacts

Figure C-14 shows the fundamental frequency distribution of T3 and T4 after LPC analysis-synthesis processing, both with and without F0 modification. F0 distribution of unprocessed speech is provided for reference. The processed(same_pitch) speech had the same F0 contour as conversational speech. Values of F0 average and standard deviation are provided in Table C.2.

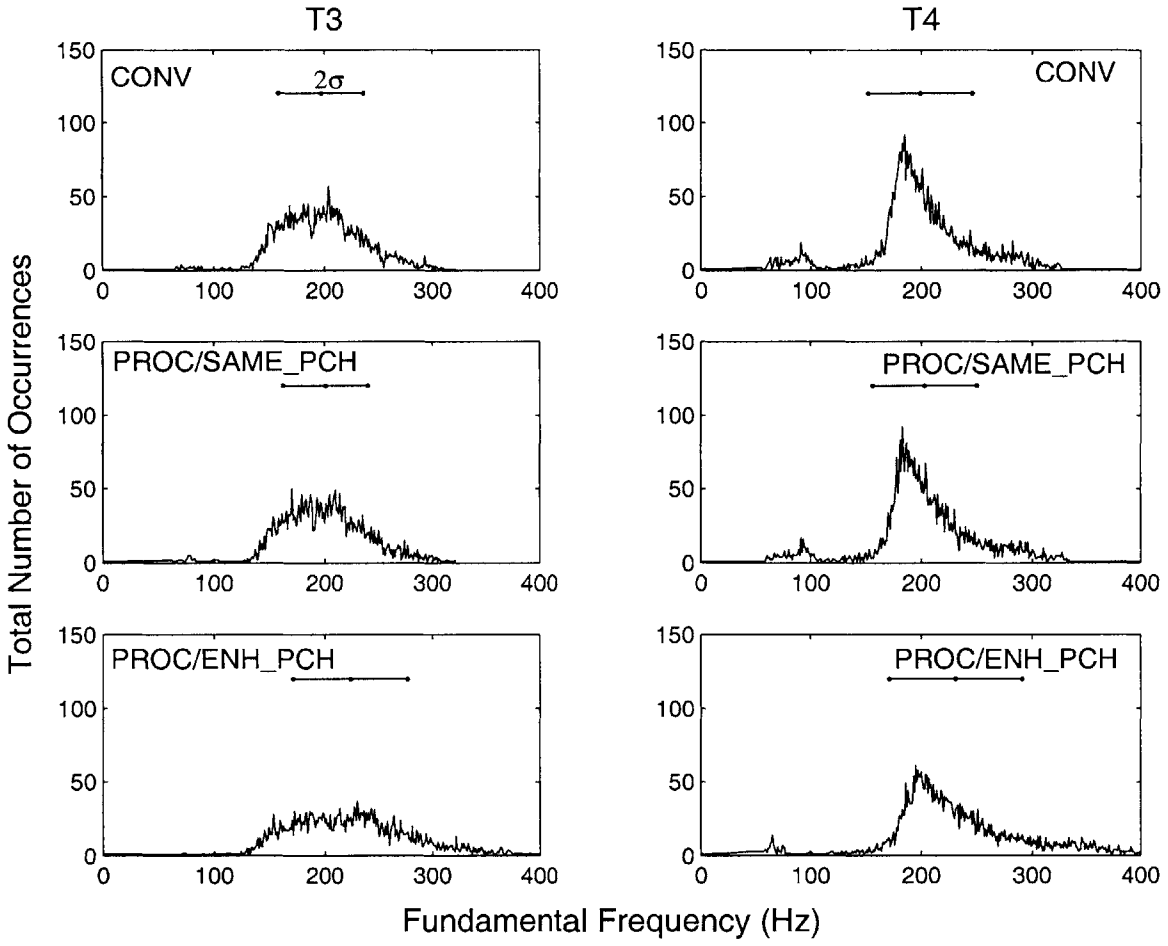

Figure C-14: Fundamental frequency distributions for conv/normal, processed(same_pitch)/normal and processed(enhanced_pitch)/normal speech of T3 and T4. Each row shows distributions for different speaking modes; columns give results for each talker.

Table C.2: Fundamental frequency distributions for T3 and T4, with and without altering the pitch via LPC processing. Conversational values before processing are provided for reference values.

Mode	T3		T4	
	Mean (Hz)	St Dev (Hz)	Mean (Hz)	St Dev (Hz)
conv/norm	197	39	199	47
proc_same_pitch/norm	201	39	203	47
proc_enh_pitch/norm	224	53	231	60

Figures C-15 and C-16 show the long-term spectra of T3 and T4 for both the twice-processed(same_formant) and the processed(enhanced_formant) conditions. From these graphs, it appears that the twice-processing scheme was successful in restoring formants to their original values.

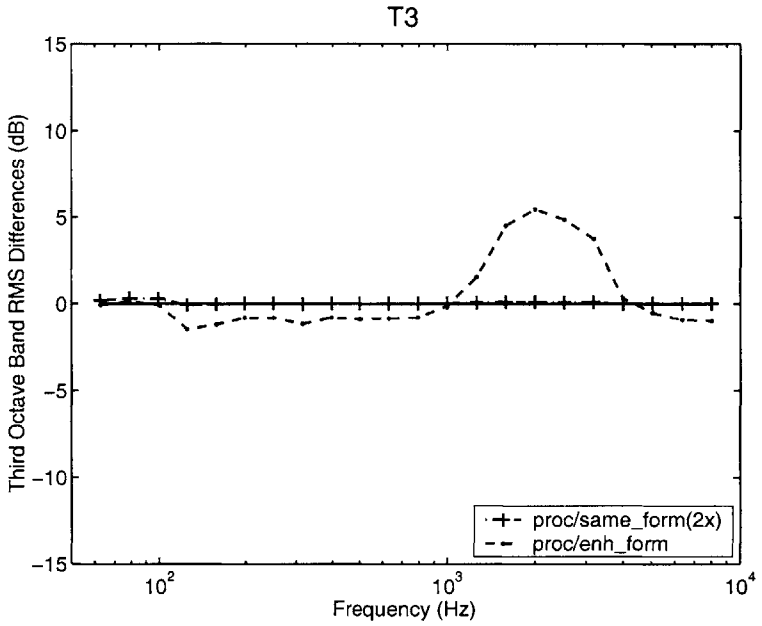

Figure C-15: Third-octave band RMS spectral differences of T3's processed(same_formant)/normal and processed(enhanced_formant)/normal modes relative to conv/normal speech.

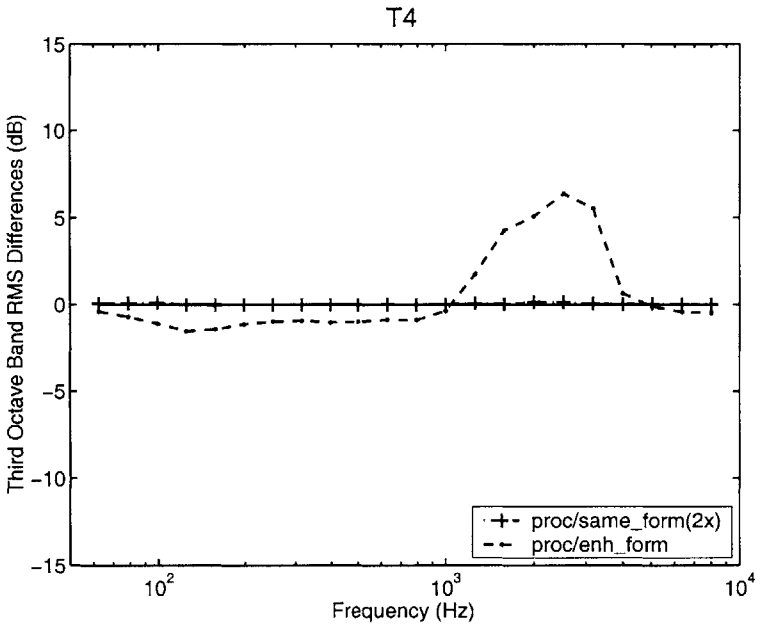

Figure C-16: Third-octave band RMS spectral differences of T4's processed(same_formant)/normal and processed(enhanced_formant)/normal modes relative to conv/normal speech.

Figures C-17 and C-18 show the intensity envelope spectra of T3 and T4 for both the twice-processed(same_envelope) and the processed(enhanced_envelope) conditions. From these graphs, it appears that the twice-processing scheme was successful in restoring envelopes to their original values.

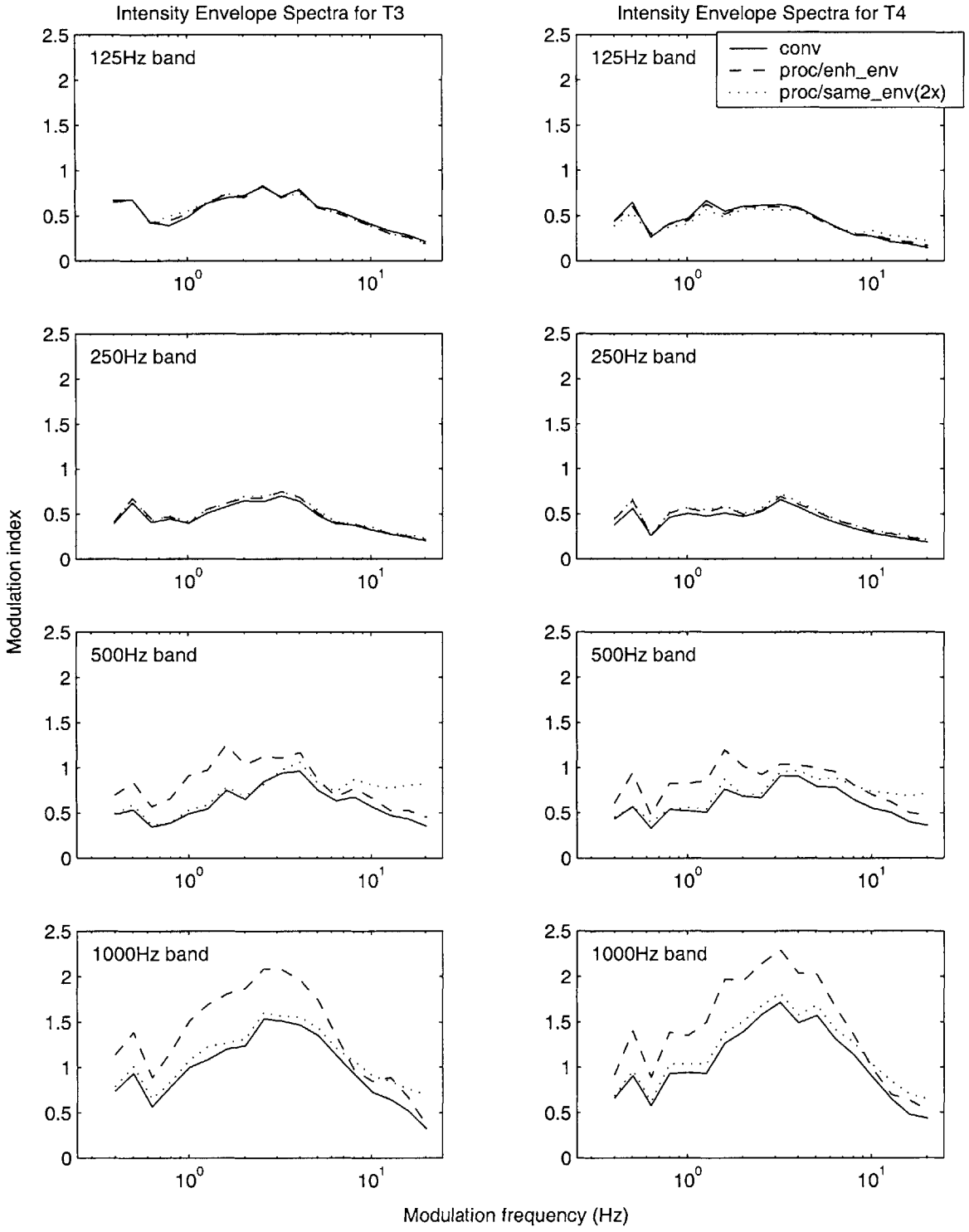

Figure C-17: Spectra of intensity envelopes for Talkers T3 and T4 in lower four octave bands for processed(same_envelope)/normal and processed(enhanced_envelope)/normal modes. Spectra for conversational envelopes is provided as a reference.

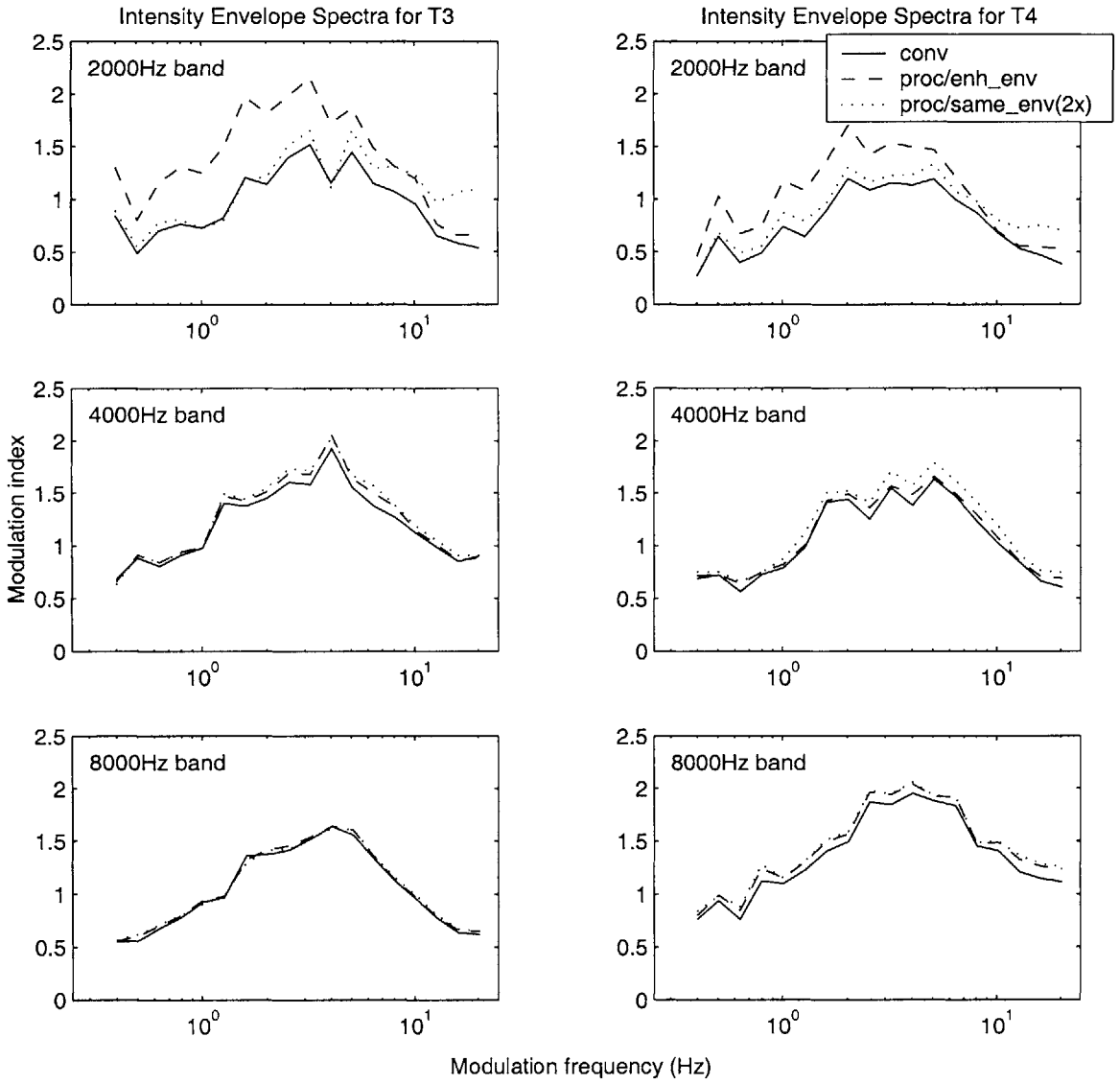

Figure C-18: Spectra of intensity envelopes for Talkers T3 and T4 in upper three octave bands for processed(same_envelope)/normal and processed(enhanced_envelope)/normal modes. Spectra for conversational envelopes is provided as a reference.

Appendix D

Listener Audiograms

The hearing levels of the five normal hearing listeners who participated in the intelligibility tests are listed in Table D.1.

Table D.1: Audiograms for the five normal-hearing listeners who participated in the intelligibility tests. Numbers reflect hearing level in dB.

Listener		250 Hz	500 Hz	1000Hz	2000 Hz	4000 Hz
L1 (DF)	R	0.0	6.0	6.0	6.0	6.0
	L	6.0	12.0	12.0	0.0	15.0
L2 (WH)	R	0.0	0.0	0.0	6.0	0.0
	L	6.0	6.0	12.0	17.5	-6.0
L3 (JM)	R	-6.0	0.0	0.0	6.0	-6.0
	L	0.0	0.0	-6.0	0.0	-6.0
L4 (CR)	R	-6.0	-6.0	-6.0	0.0	6.0
	L	6.0	12.0	6.0	0.0	6.0
L5 (TT)	R	-12.0	12.0	12.0	6.0	6.0
	L	6.0	12.0	6.0	6.0	17.5

Table D.2: Audiograms for the four normal-hearing listeners who participated in the follow-up intelligibility test. Numbers reflect hearing level in dB for the ear used in the experiment.

Listener		250 Hz	500 Hz	1000Hz	2000 Hz	4000 Hz
L6 (RB)	L	0.0	0.0	0.0	6.0	6.0
L7 (JD)	L	12.0	12.0	0.0	6.0	12.0
L8 (CH)	R	-6.0	0.0	0.0	0.0	0.0
L9 (NM)	L	0.0	0.0	0.0	0.0	6.0

Table D.3: Audiograms for the six normal-hearing listeners who participated in the final intelligibility test (T1 only). Numbers reflect hearing level in dB for the ear used in the experiment.

Listener		250 Hz	500 Hz	1000Hz	2000 Hz	4000 Hz
L10 (KF)	L	6.0	6.0	0.0	0.0	6.0
L11 (SC)	R	0.0	0.0	6.0	0.0	-6.0
L12 (FD)	L	12.0	6.0	0.0	6.0	6.0
L13 (ZS)	R	6.0	0.0	-6.0	12.0	0.0
L14 (JH)	L	0.0	6.0	-6.0	-12.0	0.0
L15 (NS)	L	6.0	12.0	-6.0	0.0	-12.0

Appendix E

Key-word Scores

This appendix presents the raw scores and corresponding percent correct key-words scores for all listeners who participated in the intelligibility experiments described in Chapter 6. A t-test was applied to percent-correct scores, after an arcsine transformation (arcsin $\sqrt{I_j/100}$) to equalize the variances, in order to determine the significance of difference between the mean of each test condition compared with conversational speech at the same speaking rate. Conditions which were significant at the 0.05 level are indicated by an asterisk.

Results for the normal hearing listeners are listed in Tables E.1 through E.4, and results for the hearing-impaired listeners are listed in Tables E.5 through E.8. A description of which sentences were used for each sentence list can be found in Appendix F.

Table E.1: Raw and percent correct key-word scores for T1.

Talker SA, List C1						
conv/norm	total	DF	WH	JM	CR	TT
cr1/list1	34	14	18	13	15	9
cr1/list2	36	11	21	19	15	6
cr1/list3	33	11	13	15	11	9
total	103	36	52	47	41	24
% correct		35.0	50.5	45.6	39.8	23.3
Talker SA, List C2						
clear/norm	total	DF	WH	JM	CR	TT
cr3/list1	37	12	30	26	19	20
cr3/list2	34	16	26	24	22	22
cr3/list3	31	14	29	25	30	23
total	102	42	85	75	71	65
% correct		41.2	***83.3**	***73.5**	69.6	***63.7**
Talker SA, List C3						
Prop C/norm	total	DF	WH	JM	CR	TT
cr1/list4	37	7	10	8	11	4
cr1/list5	35	10	16	12	10	10
cr3/list4	35	12	20	14	12	14
total	107	29	46	34	33	28
% correct		27.1	43.0	31.8	30.8	26.2
Talker SA, List C4						
Prop B/norm	total	DF	WH	JM	CR	TT
cr2/list1	36	19	26	21	23	14
cr2/list2	33	14	23	16	18	10
cr3/list5	36	19	28	21	18	16
total	105	52	77	58	59	40
% correct		***49.5**	***73.3**	55.2	***56.2**	38.1
Talker SA, List C5						
Prop A/norm	total	DF	WH	JM	CR	TT
cr2/list3	37	15	19	11	14	5
cr2/list4	36	15	20	15	14	8
cr2/list5	36	20	22	19	21	17
total	109	50	61	45	49	30
% correct		45.9	56.0	41.3	45.0	27.5

Talker SA, List C6						
Prop A+B/norm	total	DF	WH	JM	CR	TT
cr4/list4	38	24	23	23	19	15
cr4/list5	36	20	24	21	23	13
total	74	44	47	44	42	28
% correct		***59.5**	63.5	59.5	***56.8**	37.8
Talker SA, List C7						
Prop A+B+C/norm	total	DF	WH	JM	CR	TT
cr4/list1	33	10	11	9	6	9
cr4/list2	34	15	14	13	7	11
cr4/list3	38	12	15	9	8	6
total	105	37	40	31	21	26
% correct		35.2	38.1	29.5	20.0	24.8
Talker SA, List C8						
conv/slow	total	DF	WH	JM	CR	TT
sr1/list1	37	18	27	21	19	14
sr2/list2	33	12	24	18	14	14
sr3/list3	34	19	26	18	18	17
total	104	49	77	57	51	45
% correct		47.1	74.0	54.8	49.0	43.3
Talker SA, List C9						
Prop A+B/slow	total	DF	WH	JM	CR	TT
sr2/list1	35	11	10	11	10	8
sr2/list2	34	9	12	7	8	8
sr2/list3	33	19	19	17	17	14
total	102	39	41	35	35	30
% correct		38.2	40.2	34.3	34.3	29.4
Talker SA, List C10						
Prop A+B+C/slow	total	DF	WH	JM	CR	TT
sr2/list4	34	21	23	15	13	11
sr2/list5	36	29	28	22	23	18
sr1/list4	34	21	21	20	18	15
total	104	71	72	57	54	44
% correct		68.3	69.2	54.8	51.9	42.3

Table E.2: Raw and percent correct key-word scores for T2.

Talker RG, List C1						
conv/norm	total	DF	WH	JM	CR	TT
cr1/list1	35	10	18	13	12	14
cr1/list2	34	9	10	7	10	8
cr1/list3	35	15	20	15	16	10
total	104	34	48	35	38	32
% correct		32.7	46.2	33.7	36.5	30.8
Talker RG, List C2						
clear/norm	total	DF	WH	JM	CR	TT
cr3/list1	35	8	20	17	17	9
cr3/list2	33	19	20	15	20	14
cr3/list3	35	17	19	18	18	11
total	103	44	59	50	55	34
% correct		42.7	57.3	48.5	***53.4**	33.0
Talker RG, List C3						
Prop C/norm	total	DF	WH	JM	CR	TT
cr1/list4	35	7	11	6	4	6
cr1/list5	37	12	7	6	7	4
cr3/list4	34	7	6	3	5	4
total	106	26	24	15	16	14
% correct		24.5	22.6	14.2	15.7	13.2
Talker RG, List C4						
Prop B/norm	total	DF	WH	JM	CR	TT
cr2/list1	35	6	10	8	13	4
cr2/list2	34	6	10	5	11	6
cr3/list5	34	8	15	6	8	5
total	103	20	35	19	32	15
% correct		19.4	34.0	18.4	31.1	14.6
Talker RG, List C5						
Prop A/norm	total	DF	WH	JM	CR	TT
cr2/list3	35	13	14	16	12	5
cr2/list4	33	14	22	16	16	14
cr2/list5	32	15	19	14	11	9
total	100	42	55	46	39	28
% correct		42.0	55.0	46.0	40.6	28.0

Talker RG, List C6						
Prop A+B/norm	total	DF	WH	JM	CR	TT
cr4/list4	34	8	7	5	9	4
cr4/list5	36	5	9	4	4	6
total	70	13	16	9	13	10
% correct		18.6	22.9	12.9	18.6	14.3
Talker RG, List C7						
Prop A+B+C/norm	total	DF	WH	JM	CR	TT
cr4/list1	34	1	0	3	2	1
cr4/list2	36	2	5	2	2	3
cr4/list3	35	7	6	3	4	4
total	105	10	11	8	8	8
% correct		9.5	10.5	7.6	7.6	7.6
Talker RG, List C8						
conv/slow	total	DF	WH	JM	CR	TT
sr1/list1	35	12	21	16	21	15
sr2/list2	35	17	27	18	20	13
sr3/list3	34	17	22	14	17	15
total	104	46	70	48	58	43
% correct		44.2	67.3	46.2	55.8	41.3
Talker RG, List C9						
Prop A+B/slow	total	DF	WH	JM	CR	TT
sr2/list1	38	0	4	5	5	5
sr2/list2	35	5	6	5	4	5
sr2/list3	34	2	7	7	3	2
total	107	7	17	17	12	12
% correct		6.5	15.9	15.9	11.2	11.2
Talker RG, List C10						
Prop A+B+C/slow	total	DF	WH	JM	CR	TT
sr2/list4	35	19	13	15	16	10
sr2/list5	33	15	17	14	18	10
sr1/list4	36	10	14	17	17	6
total	104	44	44	46	51	26
% correct		42.3	42.3	44.2	49.0	25.0

Table E.3: Raw and percent correct key-word scores for T3.

Talker MI, List C1						
conv/norm	total	DF	WH	JM	CR	TT
cr1/list1	35	7	11	9	5	9
cr1/list2	33	9	17	6	8	9
cr1/list3	33	8	9	3	7	5
total	101	24	37	18	20	23
% correct		23.8	36.6	17.8	19.8	22.8
Talker MI, List C2						
clear/norm	total	DF	WH	JM	CR	TT
cr3/list1	34	16	18	18	17	14
cr3/list2	35	19	20	19	15	15
cr3/list3	35	17	19	10	12	11
total	104	52	57	47	44	40
% correct		***50.0**	54.8	***45.2**	***42.3**	***38.5**
Talker MI, List C3						
Prop C/norm	total	DF	WH	JM	CR	TT
cr1/list4	35	3	5	2	5	1
cr1/list5	34	3	6	3	4	3
cr3/list4	35	3	5	3	4	3
total	104	9	16	8	13	7
% correct		8.7	15.4	7.7	12.5	6.7
Talker MI, List C4						
Prop B/norm	total	DF	WH	JM	CR	TT
cr2/list1	34	7	12	5	3	6
cr2/list2	35	5	8	6	4	3
cr3/list5	34	3	13	8	6	5
total	103	15	33	19	13	14
% correct		14.6	32.0	18.4	12.6	13.6
Talker MI, List C5						
Prop A/norm	total	DF	WH	JM	CR	TT
cr2/list3	37	17	19	13	10	9
cr2/list4	34	14	19	13	14	17
cr2/list5	35	13	19	19	17	7
total	106	44	57	45	41	33
% correct		***41.5**	53.8	***42.5**	38.7	31.1

Talker MI, List C6						
Prop A+B/norm	total	DF	WH	JM	CR	TT
cr4/list4	34	7	8	8	4	2
cr4/list5	31	10	10	9	7	3
total	65	17	18	17	11	5
% correct		26.2	27.7	26.2	16.9	7.7
Talker MI, List C7						
Prop A+B+C/norm	total	DF	WH	JM	CR	TT
cr4/list1	35	2	3	3	1	1
cr4/list2	33	2	5	0	1	3
cr4/list3	36	4	3	2	3	1
total	104	8	11	5	5	5
% correct		7.7	10.6	4.8	5.0	4.8
Talker MI, List C8						
conv/slow	total	DF	WH	JM	CR	TT
sr1/list1	32	18	23	22	23	16
sr2/list2	36	19	21	23	22	18
sr3/list3	36	27	25	27	25	24
total	104	64	69	72	70	58
% correct		61.5	66.3	69.2	67.3	55.8
Talker MI, List C9						
Prop A+B/slow	total	DF	WH	JM	CR	TT
sr2/list1	37	10	11	6	6	5
sr2/list2	34	14	11	15	11	7
sr2/list3	38	10	17	8	11	12
total	109	34	39	29	28	24
% correct		31.2	35.8	26.6	25.7	22.0
Talker MI, List C10						
Prop A+B+C/slow	total	DF	WH	JM	CR	TT
sr2/list4	33	20	21	21	15	13
sr2/list5	32	19	19	15	13	11
sr1/list4	35	18	25	19	15	17
total	100	57	65	55	43	41
% correct		57.0	65.0	55.0	43.0	41.0

Table E.4: Raw and percent correct key-word scores for T4.

Talker EK, List C1						
conv/norm	total	DF	WH	JM	CR	TT
cr1/list1	35	13	18	19	13	9
cr1/list2	34	12	20	12	12	10
cr1/list3	36	16	18	14	15	11
total	105	41	56	45	40	30
% correct		39.0	53.3	42.9	38.1	28.6
Talker EK, List C11						
clear/quick	total	DF	WH	JM	CR	TT
sr2/list4	33	16	18	11	13	6
sr2/list5	34	19	18	18	13	11
sr1/list4	34	20	18	18	16	17
total	101	55	54	47	42	34
% correct		***54.5**	53.5	46.5	41.6	33.7
Talker EK, List C3						
Prop C/norm	total	DF	WH	JM	CR	TT
cr1/list4	36	3	8	6	8	3
cr1/list5	37	13	11	15	10	8
cr3/list4	37	6	12	10	11	8
total	110	22	31	31	29	19
% correct		20.0	28.2	28.2	26.4	17.3
Talker EK, List C4						
Prop B/norm	total	DF	WH	JM	CR	TT
cr2/list1	32	5	8	8	6	3
cr2/list2	35	13	13	11	11	7
cr3/list5	33	9	16	17	16	14
total	100	27	37	36	33	24
% correct		27.0	37.0	36.0	33.0	24.0
Talker EK, List C5						
Prop A/norm	total	DF	WH	JM	CR	TT
cr2/list3	36	9	18	18	12	10
cr2/list4	36	12	16	15	14	9
cr2/list5	33	13	14	11	13	12
total	105	34	48	44	39	31
% correct		32.4	45.7	41.9	37.1	29.5

Talker EK, List C6						
Prop A+B/norm	total	DF	WH	JM	CR	TT
cr4/list4	36	5	6	5	4	2
cr4/list5	34	2	3	3	0	4
total	70	7	9	8	4	6
% correct		10.0	12.9	11.4	5.7	8.6
Talker EK, List C7						
Prop A+B+C/norm	total	DF	WH	JM	CR	TT
cr4/list1	36	1	8	3	5	2
cr4/list2	37	3	3	3	2	1
cr4/list3	35	3	5	5	1	2
total	108	7	16	11	8	5
% correct		6.5	14.8	10.2	7.4	4.6
Talker EK, List C8						
conv/slow	total	DF	WH	JM	CR	TT
sr1/list1	37	20	25	25	21	18
sr2/list2	33	20	24	22	19	19
sr3/list3	35	22	21	18	19	16
total	105	62	70	65	59	53
% correct		59.0	66.7	61.9	56.2	50.5
Talker EK, List C9						
Prop A+B/slow	total	DF	WH	JM	CR	TT
sr2/list1	34	9	7	4	7	6
sr2/list2	34	12	13	10	9	12
sr2/list3	34	10	9	8	7	4
total	102	31	29	22	23	22
% correct		30.4	28.4	21.6	22.5	21.6
Talker EK, List C10						
Prop A+B+C/slow	total	DF	WH	JM	CR	TT
sr2/list4	36	16	19	17	16	11
sr2/list5	35	20	14	18	18	14
sr1/list4	34	14	15	19	15	12
total	105	50	48	54	49	37
% correct		47.6	45.7	51.4	46.7	35.2

Table E.5: Hearing-impaired listeners' raw and percent correct key-word scores for T1.

Speaker SA, List C1				
conv/norm	total	GI	RK	GT
cr1/list1	34	25	30	28
cr1/list2	36	27	28	29
cr1/list3	33	25	28	29
total	103	77	86	86
% correct		74.8	83.5	83.5
Speaker SA, List C2				
clear/norm	total	GI	RK	GT
cr3/list1	37	31	28	31
cr3/list2	34	28	22	31
cr3/list3	31	24	17	30
total	102	83	67	92
% correct		81.4	65.7	90.2
Speaker SA, List C3				
Prop C/norm	total	GI	RK	GT
cr1/list4	37	13	9	8
cr1/list5	35	15	16	22
cr3/list4	35	16	16	20
total	107	44	41	50
% correct		41.1	38.3	46.7
Speaker SA, List C4				
Prop B/norm	total	GI	RK	GT
cr2/list1	36	25	29	29
cr2/list2	33	23	17	30
cr3/list5	36	30	19	28
total	105	78	65	87
% correct		74.3	61.9	82.9
Speaker SA, List C5				
Prop A/norm	total	GI	RK	GT
cr2/list3	37	15	21	26
cr2/list4	36	16	25	33
cr2/list5	36	11	22	32
total	109	42	68	91
% correct		38.5	62.4	83.5

Speaker SA, List C6				
Prop A+B/norm	total	GI	RK	GT
cr4/list4	38	35	23	30
cr4/list5	36	33	26	31
total	74	68	49	61
% correct		***91.9**	66.2	82.4

Speaker SA, List C7				
Prop A+B+C/norm	total	GI	RK	GT
cr4/list1	33	8	9	15
cr4/list2	34	17	16	18
cr4/list3	38	16	15	13
total	105	41	40	46
% correct		39.0	38.1	43.8

Speaker SA, List C8				
conv/slow	total	GI	RK	GT
sr1/list1	37	29	25	27
sr2/list2	33	26	25	29
sr3/list3	34	22	24	31
total	104	77	74	87
% correct		74.0	71.2	83.7

Speaker SA, List C9				
Prop A+B/slow	total	GI	RK	GT
sr2/list1	35	11	10	15
sr2/list2	34	8	10	14
sr2/list3	33	16	13	18
total	102	35	33	47
% correct		34.3	32.4	46.1

Speaker SA, List C10				
Prop A+B+C/slow	total	GI	RK	GT
sr2/list4	34	22	23	29
sr2/list5	36	24	24	29
sr1/list4	34	19	30	29
total	104	65	77	87
% correct		62.5	74.0	83.7

Speaker SA, List C11				
slow/clear	total	GI	RK	GT
sr1/list5	37	31	28	32
% correct		83.8	75.7	86.5

Table E.6: Hearing-impaired listeners' raw and percent correct key-word scores for T2.

Speaker RG, List C1				
conv/norm	total	GI	RK	GT
cr1/list1	35	12	18	25
cr1/list2	34	8	15	25
cr1/list3	35	12	25	25
total	104	32	58	75
% correct		30.8	55.8	72.1
Speaker RG, List C2				
clear/norm	total	GI	RK	GT
cr3/list1	35	12	28	30
cr3/list2	33	15	27	25
cr3/list3	35	21	25	29
total	103	48	80	84
% correct		46.6	77.7	81.6
Speaker RG, List C3				
Prop C/norm	total	GI	RK	GT
cr1/list4	35	4	10	17
cr1/list5	37	4	7	14
cr3/list4	34	3	5	10
total	106	11	22	41
% correct		10.4	20.8	38.7
Speaker RG, List C4				
Prop B/norm	total	GI	RK	GT
cr2/list1	35	16	18	22
cr2/list2	34	20	17	26
cr3/list5	34	14	10	22
total	103	50	45	70
% correct		***48.5**	43.7	68.0
Speaker RG, List C5				
Prop A/norm	total	GI	RK	GT
cr2/list3	35	19	18	20
cr2/list4	33	20	23	24
cr2/list5	32	19	14	26
total	100	58	55	70
% correct		***58.0**	55.0	70.0

Speaker RG, List C6				
Prop A+B/norm	total	GI	RK	GT
cr4/list4	34	13	17	25
cr4/list5	36	11	19	19
total	70	24	36	44
% correct		34.3	51.4	62.9
Speaker RG, List C7				
Prop A+B+C/norm	total	GI	RK	GT
cr4/list1	34	2	5	3
cr4/list2	36	4	4	12
cr4/list3	35	3	7	10
total	105	9	16	25
% correct		8.6	15.2	23.8
Speaker RG, List C8				
conv/slow	total	GI	RK	GT
sr1/list1	35	21	14	27
sr2/list2	35	25	19	31
sr3/list3	34	20	20	27
total	104	66	53	85
% correct		63.5	51.0	81.7
Speaker RG, List C9				
Prop A+B/slow	total	GI	RK	GT
sr2/list1	38	6	8	11
sr2/list2	35	3	4	12
sr2/list3	34	2	7	10
total	107	11	19	33
% correct		10.3	17.8	30.8
Speaker RG, List C10				
Prop A+B+C/slow	total	GI	RK	GT
sr2/list4	35	21	18	28
sr2/list5	33	17	19	27
sr1/list4	36	19	19	22
total	104	57	56	77
% correct		54.8	53.8	74.
Speaker RG, List C11				
clear/slow	total	GI	RK	GT
sr1/list5	32	18	23	25
% correct		56.3	71.9	78.1

Table E.7: Hearing-impaired listeners' raw and percent correct key-word scores for T3.

Speaker MI, List C1				
conv/norm	total	GI	RK	GT
cr1/list1	35	13	16	20
cr1/list2	33	11	18	20
cr1/list3	33	11	14	23
total	101	35	48	63
% correct		34.7	47.5	62.4
Speaker MI, List C2				
clear/norm	total	GI	RK	GT
cr3/list1	34	19	18	26
cr3/list2	35	17	20	27
cr3/list3	35	19	15	21
total	104	55	53	74
% correct		***52.9**	51.0	71.2
Speaker MI, List C3				
Prop C/norm	total	GI	RK	GT
cr1/list4	35	7	1	8
cr1/list5	34	7	2	9
cr3/list4	35	7	6	17
total	104	21	9	34
% correct		20.2	8.7	32.7
Speaker MI, List C4				
Prop B/norm	total	GI	RK	GT
cr2/list1	34	17	13	18
cr2/list2	35	11	11	22
cr3/list5	34	13	12	26
total	103	41	36	66
% correct		39.8	35.0	64.1
Speaker MI, List C5				
Prop A/norm	total	GI	RK	GT
cr2/list3	37	19	20	28
cr2/list4	34	21	21	27
cr2/list5	35	21	15	25
total	106	61	56	80
% correct		***57.5**	52.8	***75.5**

Speaker MI, List C6				
Prop A+B/norm	total	GI	RK	GT
cr4/list4	34	15	13	19
cr4/list5	34	21	13	20
total	68	36	26	39
% correct		52.9	38.2	57.4

Speaker MI, List C7				
Prop A+B+C/norm	total	GI	RK	GT
cr4/list1	35	4	8	8
cr4/list2	33	7	5	7
cr4/list3	36	5	3	6
total	104	16	16	21
% correct		15.4	15.4	20.2

Speaker MI, List C8				
conv/slow	total	GI	RK	GT
sr1/list1	32	22	18	25
sr2/list2	36	29	30	31
sr3/list3	36	30	26	33
total	104	81	74	89
% correct		77.9	71.2	85.6

Speaker MI, List C9				
Prop A+B/slow	total	GI	RK	GT
sr2/list1	37	10	7	12
sr2/list2	34	14	8	13
sr2/list3	38	13	8	7
total	109	37	23	32
% correct		33.9	21.1	29.4

Speaker MI, List C10				
Prop A+B+C/slow	total	GI	RK	GT
sr2/list4	33	17	20	19
sr2/list5	32	18	18	19
sr1/list4	35	21	26	29
total	100	56	64	67
% correct		56.0	64.0	67.0

Speaker MI, List C11				
clear/slow	total	GI	RK	GT
sr1/list5	35	25	23	23
% correct		71.4	65.7	65.7

Table E.8: Hearing-impaired listeners' raw and percent correct key-word scores for T4.

Speaker EK, List C1				
conv/norm	total	GI	RK	GT
cr1/list1	35	21	25	29
cr1/list2	34	17	24	25
cr1/list3	36	20	24	29
total	105	58	73	83
% correct		55.2	69.5	79.0
Speaker EK, List C2				
clear/norm	total	GI	RK	GT
qr1/list1	33	22	17	27
qr1/list2	34	21	23	27
qr1/list3	34	19	24	28
total	101	62	64	82
% correct		61.4	63.4	81.2
Speaker EK, List C3				
Prop C/norm	total	GI	RK	GT
cr1/list4	36	4	11	11
cr1/list5	37	6	9	18
cr3/list4	37	2	8	13
total	110	12	28	42
% correct		10.9	25.5	38.2
Speaker EK, List C4				
Prop B/norm	total	GI	RK	GT
cr2/list1	35	7	10	15
cr2/list2	35	14	10	18
cr3/list5	33	20	18	22
total	103	41	38	55
% correct		39.8	36.9	53.4
Speaker EK, List C5				
Prop A/norm	total	GI	RK	GT
cr2/list3	36	14	24	23
cr2/list4	36	21	22	26
cr2/list5	33	19	20	24
total	105	54	66	73
% correct		51.4	62.9	69.5

Speaker EK, List C6				
Prop A+B/norm	total	GI	RK	GT
cr4/list4	36	2	13	14
cr4/list5	34	1	12	15
total	70	3	25	29
% correct		4.3	35.7	41.4

Speaker EK, List C7				
Prop A+B+C/norm	total	GI	RK	GT
cr4/list1	36	4	9	9
cr4/list2	37	6	4	8
cr4/list3	35	4	3	10
total	108	14	16	27
% correct		13.0	14.8	25.0

Speaker EK, List C8				
conv/slow	total	GI	RK	GT
sr1/list1	37	10	23	29
sr2/list2	33	16	28	27
sr3/list3	35	21	17	27
total	105	47	68	83
% correct		44.8	64.8	79.0

Speaker EK, List C9				
Prop A+B/slow	total	GI	RK	GT
sr2/list1	34	6	10	15
sr2/list2	34	10	6	15
sr2/list3	34	5	13	16
total	102	21	29	46
% correct		20.6	28.4	45.1

Speaker EK, List C10				
Prop A+B+C/slow	total	GI	RK	GT
sr2/list4	36	21	22	27
sr2/list5	35	21	22	23
sr1/list4	34	26	23	27
total	105	68	67	77
% correct		64.7	63.8	73.3

Speaker EK, List C11				
clear/slow	total	GI	RK	GT
sr1/list5	34	16	29	28
% correct		47.1	85.3	82.4

Appendix F

Sentence Lists

All sentences used in intelligibility tests were from the corpus of sentences described by Picheny *et al.*[44]. Tables F.1 and F.2 explains which sentences were used in each of the sentence lists.

Table F.1: Sentence lists recorded by T1 and T2 for formal intelligibility tests. SP, LST, and SUB correspond to Picheny's notation for describing the corpus in Appendix B of his thesis[44].

List	Sublist	Talker	
		T1	T2
C1	cr1/list1	SP1/LST1/SUB1	SP3/LST1/SUB1
	cr1/list2	SP1/LST1/SUB2	SP3/LST1/SUB2
	cr1/list3	SP1/LST1/SUB3	SP3/LST1/SUB3
C2	cr3/list1	SP1/LST3/SUB1	SP3/LST3/SUB1
	cr3/list2	SP1/LST3/SUB2	SP3/LST3/SUB2
	cr3/list3	SP1/LST3/SUB3	SP3/LST3/SUB3
C3	cr1/list4	SP1/LST1/SUB4	SP3/LST1/SUB4
	cr1/list5	SP1/LST1/SUB5	SP3/LST1/SUB5
	cr3/list4	SP1/LST3/SUB4	SP3/LST3/SUB4
C4	cr2/list1	SP1/LST2/SUB1	SP3/LST2/SUB1
	cr2/list2	SP1/LST2/SUB2	SP3/LST2/SUB2
	cr3/list5	SP1/LST3/SUB5	SP3/LST3/SUB5
C5	cr2/list3	SP1/LST2/SUB3	SP3/LST2/SUB3
	cr2/list4	SP1/LST2/SUB4	SP3/LST2/SUB4
	cr2/list5	SP1/LST2/SUB5	SP3/LST2/SUB5
C6	cr4/list4	SP1/LST4/SUB4	SP3/LST4/SUB4
	cr4/list5	SP1/LST4/SUB5	SP3/LST4/SUB5
C7	cr4/list1	SP1/LST4/SUB1	SP3/LST4/SUB1
	cr4/list2	SP1/LST4/SUB2	SP3/LST4/SUB2
	cr4/list3	SP1/LST4/SUB3	SP3/LST4/SUB3
C8	sr1/list1	SP1/LST6/SUB1	SP3/LST6/SUB1
	sr1/list2	SP1/LST6/SUB2	SP3/LST6/SUB2
	sr1/list3	SP1/LST6/SUB3	SP3/LST6/SUB3
C9	sr2/list1	SP1/LST7/SUB1	SP3/LST7/SUB1
	sr2/list2	SP1/LST7/SUB2	SP3/LST7/SUB2
	sr2/list3	SP1/LST7/SUB3	SP3/LST7/SUB3
C10	sr2/list4	SP1/LST7/SUB4	SP3/LST7/SUB4
	sr2/list5	SP1/LST7/SUB5	SP3/LST7/SUB5
	sr1/list4	SP1/LST6/SUB4	SP3/LST6/SUB4

Table F.2: Sentence lists recorded by T3 and T4 for formal intelligibility tests. SP/LST/and SUB correspond to Picheny's notation for describing the corpus in Appendix B of his thesis[44].

List	Sublist	Talker	
		T3	T4
C1	cr1/list1	SP2/LST8/SUB1	SP1/LST8/SUB1
	cr1/list2	SP2/LST8/SUB2	SP1/LST8/SUB2
	cr1/list3	SP2/LST8/SUB3	SP1/LST8/SUB3
C2	cr3/list1	SP2/LST10/SUB1	N/A
	cr3/list2	SP2/LST10/SUB2	N/A
	cr3/list3	SP2/LST10/SUB3	N/A
C3	cr1/list4	SP2/LST8/SUB4	SP1/LST8/SUB4
	cr1/list5	SP2/LST8/SUB5	SP1/LST8/SUB5
	cr3/list4	SP2/LST10/SUB4	SP1/LST10/SUB4
C4	cr2/list1	SP2/LST9/SUB1	SP1/LST9/SUB1
	cr2/list2	SP2/LST9/SUB2	SP1/LST9/SUB2
	cr3/list5	SP2/LST10/SUB5	SP1/LST10/SUB5
C5	cr2/list3	SP2/LST9/SUB3	SP1/LST9/SUB3
	cr2/list4	SP2/LST9/SUB4	SP1/LST9/SUB4
	cr2/list5	SP2/LST9/SUB5	SP1/LST9/SUB5
C6	cr4/list4	SP2/LST11/SUB4	SP1/LST11/SUB4
	cr4/list5	SP2/LST11/SUB5	SP1/LST11/SUB5
C7	cr4/list1	SP2/LST11/SUB1	SP1/LST11/SUB1
	cr4/list2	SP2/LST11/SUB2	SP1/LST11/SUB2
	cr4/list3	SP2/LST11/SUB3	SP1/LST11/SUB3
C8	sr1/list1	SP2/LST13/SUB1	SP1/LST13/SUB1
	sr1/list2	SP2/LST13/SUB2	SP1/LST13/SUB2
	sr1/list3	SP2/LST13/SUB3	SP1/LST13/SUB3
C9	sr2/list1	SP2/LST14/SUB1	SP1/LST14/SUB1
	sr2/list2	SP2/LST14/SUB2	SP1/LST14/SUB2
	sr2/list3	SP2/LST14/SUB3	SP1/LST14/SUB3
C10	sr2/list4	SP2/LST14/SUB4	SP1/LST14/SUB4
	sr2/list5	SP2/LST14/SUB5	SP1/LST14/SUB5
	sr1/list4	SP2/LST13/SUB4	SP1/LST13/SUB4
C11	qr1/list1	N/A	SP1/LST12/SUB1
	qr1/list2	N/A	SP1/LST12/SUB2
	qr1/list3	N/A	SP1/LST12/SUB3

Appendix G

Phonetic Labels

Tables G.1 lists the phonetic labels used in this thesis and provides example words for clarification when necessary.

Table G.1: Pronunciation guide for phonetic labels.

Phone	Description	Example Word	Phone	Description	Example Word
aa		hot	jh		june
ae		van	k		
ah		but	l		
ao		bought	m		
aw		how	n		
ax	schwa	about	ng		
ay		bite	ow		boat
b			oy		toy
bst	stop burst		p		
ch			r		
cl	closure		s		
d			sh		
dh	voiced "th"	then	sil	silence	
dx	flap	muddy, dirty	t		
eh		bet	th	unvoiced "th"	thin
el	syllabic "l"	bottle	em	syllabic "m"	bottom
en	syllabic "n"	button	uh		hood
er			uw		shampoo
ey		bait	v		
f			vcl	voiced closure	
g			w		
h			y		
hw	aspirated "w"		z		
ih		bit	zh		
iy		beet			

Bibliography

[1] S. Arlinger and H. A. Gustafsson. Masking of speech by amplitude-modulated noise. *Journal of Sound and Vibration*, 15(3):441–445, 1991.

[2] Carl A. Bennett and Norman L. Franklin. *Statistical Analysis in Chemistry and the Chemical Industry*, chapter 7, pages 319–468, 708–709. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, Inc., London, 1954.

[3] A.R. Bradlow, G.M. Toretta, and D.B. Pisoni. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3–4):225–272, December 1996.

[4] H. T. Bunnell. On enhancement of spectral contrast in speech for hearing-impaired listeners. *Journal of the Acoustical Society of America*, 88(6):2546–2556, December 1990.

[5] D. Byrne and H. Dillon. The national acoustic laboratories new procedure for selecting the gain and frequency response of a hearing aid. *Ear and Hearing*, 7:257–265, 1986.

[6] F. R. Chen. Acoustic characteristics and intelligibility of clear and conversational speech. Master's project, Massachusetts Institute of Technology, Research Laboratory of Electronics, May 1980.

[7] S. Choi. The effect of pauses on the intelligibility of sentences. Bachelor's project, Massachusetts Institute of Technology, Research Laboratory of Electronics, September 1987.

[8] P. Denes. Effect of duration on perception of voicing. *J. Acoust. Soc. Am.*, 27(4):761–764, 1955.

[9] A.K. Dix. *personal communication*, 1996.

[10] R. Drullman. Temporal envelope and fine structure cues for speech intelligibility. *Journal of the Acoustical Society of America*, 97(1):585–592, January 1995.

[11] R. Drullman, J. M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America*, 95(5):2670–2680, April 1994.

[12] R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95(2):1053–1064, February 1994.

[13] Thomas J. Edwards. Multiple features analysis of intervocalic english plosives. *J. Acoust. Soc. Am.*, 69(2):535–547, February 1981.

[14] Gene V Glass and Kenneth D. Hopkins. *Statistical Methods in Education and Psychology.* Academic Press, New York, New York, January 1968.

[15] S. Gordon-Salant. Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects. *Journal of the Acoustical Society of America*, 81(4):1199–1202, April 1987.

[16] K.W. Grant and L.D. Braida. Articulation index for auditory-visual input. *Journal of the Acoustical Society of America*, 89(6):2952–2960, June 1991.

[17] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32:236–243, 1984.

[18] S. E. Hargus and S. Gordon-Salant. Accuracy of speech intelligibility index predictions for noise-masked young listeners with normal hearing and for elderly listeners with hearing impairment. *Journal of Speech and Hearing Research*, 38:234–243, February 1995.

[19] H. Hawkins and S.S. Stevens. The masking of pure tones and of speech by white noise". *J. Acoust. Soc. Am.*, 22(1):6–13, January 1950.

[20] J. Hillenbrand, L.A. Getty, M.J. Clark, and K. Wheeler. Acoustic characteristics of american english vowels. *J. Acoust. Soc. Am.*, 97(5):3099–3111, 1995.

[21] J.M. Hillenbrand and T.M. Nearey. Identification of resynthesized /hvd/ utterances: Effects of formant contour. *J. Acoust. Soc. Am.*, 105(6):3509–3523, 1999.

[22] T. Houtgast, H. Steeneken, and R. Plomp. Predicting speech intelligibility in rooms from the modulation transfer function. *Acustica*, 46(1):60–72, 1980.

[23] T. Houtgast and H.J.M. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, 77(3):1069–1077, March 1985.

[24] J.E. Huber, E.T. Stathopoulos, G.M. Curione, T.A. Ash, and K. Johnson. Formants of children, women, and men: The effects of vocal intensity variation. *J. Acoust. Soc. Am.*, 106(3):1532–1542, 1999.

[25] L.E. Humes, D.D. Dirks, T.S. Bell, C. Ahlstrom, and G.E. Kincaid. Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners. *J. Speech Hear. Res.*, 29:447–462, 1986.

[26] M. Kahn and P. Garst. The effects of five voice characteristics on lpc quality. *Proceedings of IEEE Int. Conf. on Acoust., Sp., and Sig. Proc.*, 2:531–534, 1983.

[27] J. M. Kates. Speech enhancement based on a sinusoidal model. *Journal of Speech and Hearing Research*, 37:449–464, April 1994.

[28] J. C. Krause. The effects of speaking rate and speaking mode on intelligibility. Master's project, Massachusetts Institute of Technology, Research Laboratory of Electronics, August 1995.

[29] J.C. Krause and L.D. Braida. Properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America*, 100:S2828, 1996.

[30] K.D. Kryter and J.C.R. Licklider. Premodulation clipping in am voice communication. *Journal of the Acoustical Society of America*, 19(1):125–131, 1947.

[31] P. Ladefoged. *A Course in Phonetics.* Harcourt Brace Jovanovich, Inc., New York, New York, 1975.

[32] J.C.R. Licklider. Effects of amplitude distortion on the intelligibility of speech. *Journal of the Acoustical Society of America*, 29(12):429–434, 1946.

[33] J.C.R. Licklider, M.E. Hawley, and R.A. Walking. Influences of variations in speech intensity and other factors upon the speech spectrum. *Journal of the Acoustical Society of America*, 27:207, 1955.

[34] Qing-Guang Liu, Benoit Champagne, and Peter Kabal. A microphone array processing technique for speech enhancement in a reverberant space. *Speech Communication*, 18:317–334, 1996.

[35] D. Malah. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):121–133, 1979.

[36] G. A. Miller and P. E. Nicely. An analysis of perceptual confusions among some english consonants. *Journal of the Acoustical Society of America*, 27(2):338–352, March 1955.

[37] A. A. Montgomery and R. A. Edge. Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults. *Journal of Speech and Hearing Research*, 31:386–393, September 1988.

[38] R.J. Niederjohn and J.H. Grotelueschen. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(4):277–282, 1976.

[39] Michael Nilsson, Sigfrid D. Soli, and Jean A. Sullivan. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, 95:1085–1099, February 1994.

[40] M.T. Ochs, L.E. Humes, R.N. Ohde, and D.W. Grantham. Frequency discrimination ability and stop-consonant identification in normally hearing and hearing-impaired subjects. *Journal of Speech and Hearing Research*, 32(1):133–142, 1989.

[41] K. L. Payton and L. D. Braida. Determining the speech transmission index directly from speech waveforms. *submitted to Journal of the Acoustical Society of America*, 1997.

[42] K. L. Payton, R. M. Uchanski, and L. D. Braida. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America*, 95(3):1581–1592, March 1994.

[43] G.E. Peterson and H.L. Barney. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, 24:175–184, 1952.

[44] M. A. Picheny. *Speaking Clearly for the Hard of Hearing.* PhD dissertation, Massachusetts Institute of Technology, Research Laboratory of Electronics, June 1981.

[45] M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, 28:96–103, March 1985.

[46] M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29:434–446, December 1986.

[47] M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research*, 32:600–603, September 1989.

[48] J.M. Pickett. Effects of vocal force on the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, 28:902–905, 1956.

[49] M. R. Schroeder. Reference signal for signal quality studies. *Journal of the Acoustical Society of America*, 44:1735–1736, 1968.

[50] R. V. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270:303–304, October 1995.

[51] P. Tallal, S.L. Miller, G. Bedi, G. Byma, X. Wang, S.S. Nagarajan, C. Schreiner, W.M. Jenkins, and M.M. Merzenich. Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science*, 271:81–84, January 1996.

[52] P. Tallal and M. Piercy. Developmental aphasia: The perception of brief vowels and extended stop consonants. *Neuropsychologia*, 13:69–74, 1975.

[53] C. W. Turner, S. J. Smith, P. L. Aldridge, and S. L. Stewart. Formant transition duration and speech recognition in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 101(5):2822–2825, May 1997.

[54] R. M. Uchanski. *Spectral and Temporal Contributions to Speech Clarity for Hearing Impaired Listeners.* PhD dissertation, Massachusetts Institute of Technology, Research Laboratory of Electronics, May 1988.

[55] R. M. Uchanski, S. Choi, L. D. Braida, C. M. Reed, and N. I. Durlach. Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research*, 39(3):494–509, June 1996.