An Investigation of Clearly Spoken Speech and Possibilities of Intelligibility Enhancement by Redistribution of Energy

by

Kenneth Thomas Schutte

B.S. E.E. University of Illinois at Urbana-Champaign (2001)

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

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2003

Author .....
Department of Electrical Engineering and Computer Science
May 9, 2003

Certified by ....
Jae S. Lim
Professor
Thesis Supervisor

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

An Investigation of Clearly Spoken Speech and Possibilities of Intelligibility Enhancement by Redistribution of Energy

by

Kenneth Thomas Schutte

Submitted to the Department of Electrical Engineering and Computer Science
on May 9, 2003, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Several recent studies have examined the differences between speech spoken in a conversational manner and speech spoken in a highly intelligible manner, i.e. clear speech. If these differences can be isolated and quantified, it may be possible to design a signal processing algorithm able to increase the intelligibility of conversational speech by automatically introducing the important properties of clear speech.

Specific examples of these two speaking modes are examined to precisely identify where and why intelligibility improvements occur in clear speech. It is shown that a majority of these improvements may be due to a small number of segments in clear speech. While several acoustic measurements are examined to explain why these intelligibility differences occur, the analysis concentrates on segmental energy and power levels. The intelligibility of conversational speech is shown to drastically decrease throughout the course of a sentence, and it is suggested that this is caused primarily by the speaker trailing-off in power level. Clear speech exhibits similar trends, but to much less of an extent. Also, clear speech utterances contain relatively more of their energy in keywords, thereby increasing the signal-to-noise ratio during portions of the sentence crucial to intelligibility.

Several conversational utterances are modified to set segmental power levels equal to those of the corresponding segments in the clear speech utterance. Preliminary testing indicates that approximately one-third of the intelligibility gap between clear and conversational speech can be attributed to how the energy of a given sentence is distributed amongst its phonetic segments.

Thesis Supervisor: Jae S. Lim
Title: Professor of Electrical Engineering

Acknowledgments

I would like to thank Professor Lim for his support over the past two years. His guidance on this thesis and his teaching in the classroom have taught me a great deal about both the technical and non-technical sides of research in EECS and signal processing.

Also, I must thank the other members of the Advanced Television and Signal Processing group: Brian, James, and Cindy. Their constant willingness to go out of their way to help has been greatly appreciated.

A special thanks goes to Jeannie Krause for offering advice several times through the course of this project as well as providing the data and previous research which made this thesis possible.

Thank you to the five volunteers who participated in the listening experiments.

Most importantly, I must thank my parents for their continuous encouragement, motivation, and support throughout my life. I could not have made it this far without them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When speaking in the presence of background noise or when speaking to the hearing impaired, a conversational manner of speech may not effectively convey the desired message. In these difficult environments, humans tend to change their style of speech in order to be better understood. This altered style, referred to as clear speech, has been shown to have increased intelligibility over conversational speech when presented to listeners in presence of various noise environments. The intelligibility advantages of clear speech tend to be very robust and have been shown to be independent of listener, presentation level, and frequency-gain characteristic [6].

Recently, researchers have attempted to measure acoustical differences between clear and conversational speech to identify which characteristics of clear speech lead to its increased intelligibility [3, 7]. If such characteristics are successfully isolated and quantified, it may be possible to develop an algorithm which transforms conversational speech into some approximation of clear speech. Such a transformation would have many potential applications, including hearing aid improvements. Despite some interesting findings, the specific properties of clear speech leading to its high intelligibility have yet to be pinpointed.

This thesis is another attempt at analyzing clear speech to determine the causes of its intelligibility advantages. However, a slightly different approach will be taken. Results from intelligibility tests will be used to isolate where the advantages of clear speech actually occur and to select a few example utterances for further analysis. The

selected conversational utterances will be modified "by-hand" to more closely resemble clear speech, and effects of this modification on intelligibility will be measured. Using transformations tailored for each of a few select utterances allows testing the effects of some desired property without requiring a general rule of how this property changes in clear speech. If successful, this would not provide a general speech enhancement algorithm, but it would offer valuable insight into how one might be devised.

The rest of this thesis is organized as follows. Chapter 2 reviews major results from previous studies on this topic. Chapter 3 investigates where intelligibility increases occur in clear speech and proposes a possible partial explanation of these improvements. Based on this explanation, Chapter 4 performs a simple modification of conversational speech and tests its effect on intelligibility. Finally, Chapter 5 contains a discussion of the results and a conclusion.

Chapter 2

Background

In an attempt to automatically introduce properties of clear speech into conversational speech, one must first explore the quantitative differences between these two speaking modes. Much of this thesis relies upon several recent studies which set out to accomplish this task. A series of four papers by Picheny, Durlach, Braida, and Uchanski entitled "Speaking Clearly for the Hard of Hearing" [6, 7, 8, 12] analyzed many properties of clear speech, particularly speaking rate. Two theses by Krause [2, 3] conducted follow-up studies on clear speech at normal rates. This chapter provides a brief summary of their research and their conclusions.

An initial study [6] was conducted to more precisely measure the benefits of clear speech over conversational speech. Nonsense sentences spoken in both modes were presented to five hearing impaired listeners to compare intelligibility between the two. The clear speech was shown to be consistently more intelligible than conversational speech. The percentage of keywords (nouns, adjectives, and adverbs) correctly identified in clear speech was, on average, 17 points higher than that of conversational. The intelligibility increase was similar across speakers, listeners, presentation levels, and frequency dependent linear amplification. Additional testing showed that these benefits also extend to normal hearing listeners in the presence of wideband noise [5].

With the intelligibility benefits of clear speech established, measurements were taken to identify features of clear speech that could cause these improvements. While a variety of these measurements are discussed later in the chapter, the property of

speaking rate will be singled out for initial consideration in the following section. Speaking rate is the most drastic and consistent difference between clear and conversational speech; therefore, much research has been focused on its role in speech intelligibility.

2.1 The Role of Speaking Rate

Clear speech has been measured to be around 90-100 words/min, roughly half the rate of conversational's 160-200 words/min [6]. This reduced rate is accomplished both by increasing the duration and frequency of pauses and by increasing the duration of speech sounds. In contrast, when a talker is just asked to speak slowly, the majority of the rate reduction is due to increases in pause length.

Picheny *et al.* performed several studies intended to determine the interaction between speaking rate and intelligibility. In one experiment, time-scale modification was used to transform segments in clear speech so that their durations matched those of the corresponding segments in conversational speech. Also, conversational speech was modified to have the speaking rate characteristics of clear speech. Tests revealed that the non-uniform time-expansion of conversational speech slightly reduced intelligibility. Also, the non-uniform time-compression of clear speech was less intelligible than the original conversational speech [12]. These results suggest that speaking rate may not be a primary cause of the high intelligibility of clear speech.

Until recently, there was no solid evidence of clear speech at normal rates (either natural or synthetic). However, in a study by Krause [4], experienced speakers were trained to produce speech at three different rates (slow, normal, and quick) in each of two modes (clear and conversational)[1]. Tests revealed that clear/slow had an 18 point improvement in intelligibility, which is similar to past scores of clear speech. More importantly, clear/norm was 14 points more intelligible than conv/norm. This implies that properties other than overall speaking rate are responsible for the high

[1]These various combinations will be referred to by *mode/rate* (i.e. clear/slow, conv/norm, clear/norm, etc). "conv" will be used as short for conversational, and "norm" short for normal rate.

intelligibility of clear/norm.

While speakers in Krause's study were successfully trained to produce clear speech at normal rates, it appears that this is not a trivial task. In a previous study [12], an accomplished professional speaker received less elaborate training and his attempt at clear/norm speech did not show significant intelligibility improvements over his conv/norm speech. Also, Krause's study put much effort into carefully choosing and training the speakers; of 15 originally chosen, only two ended up being accepted for the analysis. Therefore, producing clear speech at normal rates is not an easy task. Only some portion of speakers may be able to produce it naturally, and it may require extensive training for those that can.

2.2 Acoustical Analysis

This section summarizes findings from previous studies on the characteristics (other than speaking rate) that distinguish conversational from clear speech. The first sub-section will discuss results from [7], which was an analysis of clear speech with significantly reduced speaking rates. The following sub-section will compare and contrast those results with those from [3], which dealt with clear speech at normal rates.

2.2.1 Clear Speech

In the second paper of their series [7], Picheny *et al.* performed a detailed analysis of acoustical characteristics of clear and conversational speech in an effort to determine what characteristics were essential to the high intelligibility of clear speech. This analysis was broadly divided into three levels of detail: global, phonological, and phonetic.

Differences in global properties occur at the sentence level and include such factors as speaking rate, fundamental frequency (pitch), pause distribution, and long-term spectra. There was a slight tendency for higher fundamental frequency and larger range in fundamental frequency in clear speech. Differences in long term spectra were small. Aside from speaking rate, no global phenomena were considered significantly

different between the two modes.

The second broad class of transformations, characterized as phonological, includes insertions, deletions and feature changes of phonemes. These are classified into six categories: vowel modification, burst elimination, sound insertion, degemination, alveolar flap, and sound deletion. Vowel modification or reduction was about half as common in clear speech as in conversational speech. Burst elimination was much more common in conversational speech, especially for word-final stop consonants. Sound insertion (inserting a schwa vowel after a voiced consonant) was observed in clear speech. These results tend to match intuition on properties of clearly enunciated speech. However, the extent to which these phonological transformations are necessary for high intelligibility is unknown.

At the closest level of analysis are the phonetic phenomena, which classify amplitudes, spectra, and durations of individual phones and phone classes. Studies of phonetic level phenomena have included segmental power, segmental phone duration, short-term RMS spectra, vowel formant frequencies, consonant-vowel ratio (CVR), and voice-onset time. Of these various measurements, the only significant effect noted by Picheny *et al.* was a tendency for clear speech to have increased RMS intensities for obstruents and stop consonants [7].

2.2.2 Clear Speech at Normal Rates

Krause performed a similar analysis on the database of clear/norm speech to determine if characteristics of clear speech were different when spoken at normal rates [3]. This analysis consisted of data collected from two speakers: one male and one female. Again, results are divided into global, phonological, and phonetic categories.

Spoken at normal rates, clear speech had several global characteristics that differed from clear/slow. Clear/norm had similar pause length distributions to conv/norm due to the constraint that sentences have similar durations. Also, it was concluded that long-term spectra of clear/norm had relatively more energy above 1 kHz than conv/norm. While clear speech was generally spoken louder than conversational and speaking loudly can result in an increase of high frequency energy, it was concluded

that level differences were not large enough to be responsible for the long-term spectral differences between clear/norm and conv/norm.

Fundamental frequencies tended to be higher and have larger range for clear/norm speech when compared to conv/norm. However this was much more pronounced in the male speaker than the female speaker. Investigation of temporal envelope modulations revealed an increase in modulation depth for low frequencies (<3-4 Hz) in clear speech. This effect was more pronounced in clear/slow, but was still present in clear/norm, especially in the male speaker, and thus it was included as a possible important feature of clear/norm.

Phonological differences between conv/norm and clear/norm were much less drastic than they were between conv/norm and clear/slow. The female speaker showed a greater tendency to release stop bursts in clear/norm, but the male did not. Other differences in phonological measurements were considered inconsequential or the data was deemed inconclusive.

Statistical analysis of 43 phones was conducted to measure various phonetic phenomena. Both the relative power of phonetic segments and consonant-vowel ratio (CVR) were shown not to be a necessary condition for highly intelligible speech[2]. While the average duration of some phones differed in clear/norm and conv/norm, they were increased for one speaker and decreased for the other (on average). Opposite trends between speakers were also observed in voice-onset time (VOT) calculations. Such observations suggest that the two speakers used different strategies to produce clear speech.

In short-term spectra, nearly all clear/norm vowels had relatively higher spectral prominences at formant frequencies in comparison to conv/norm. It had been previously hypothesized that formant frequencies appear closer to their target values and cluster more closely in clear speech [1]. However, these claims were not supported by this data. The data showed no significant change in vowel formant frequencies for clear/norm speech. However, the formant bandwidths were shown to be somewhat

[2]Other studies [1] concluded that CVR could play a pivotal role in the ability to correctly identify isolated syllables.

narrower in clear speech.

In summary, Table 2.1 lists properties which differed significantly between clear/norm and conv/norm. Notice that many of these differences are speaker-dependent. This suggests that different speakers may utilize different strategies for eliciting clear speech at normal rates. Some issues relating to this will be discussed throughout the thesis.

Table 2.1: Properties of clear/norm in relation to conv/norm for the two speakers studied. Table reproduced from [3].

Property	**Difference**	**Male speaker**	**Female speaker**
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)
Fundamental Frequency	Greater average, range	Yes	No
Word-initial stops	Increased VOT	Yes	No
Word-final stops	Bursts released more often	No	Yes

2.3 Attempted Transformations of Conversational Speech

As previously stated, a major objective for quantifying the acoustical differences between clear and conversational speech is to be able to apply that knowledge to a system for automatic intelligibility improvement. Krause attempted this using a set of transformations made to introduce characteristics of clear/norm speech into conv/norm [3]. Three modifications were developed based on selected properties from Table 2.1. The properties chosen as significant to increased intelligibility were pitch, energy in vowel formant frequencies, and temporal envelope modulations.

The first modification consisted of weighting the spectrum around F_2 and F_3 during voiced sections of speech. This was done by multiplying the short-time fourier

transform (STFT) magnitude by a hamming window (centered between F_2 and F_3) during times of voicing and using an iterative method to estimate a time-domain waveform from the new STFT magnitude. The second modification used an LPC method to modify the fundamental frequencies to have a higher mean and larger range. Due to their different behaviors in pitch, this transformation was designed to be slightly different for the two speakers. The third modification increased modulation depth for low frequencies (3-4 Hz) in the intensity envelopes of several octave bands.

Intelligibility tests were then conducted to measure the effectiveness of these three modifications. Listeners were presented with a total of 10 different modes of speech including conv/slow, conv/norm, clear/norm, and conv/norm processed with various combinations of the three described modifications. For normal hearing listeners, only clear/norm and conv/norm with the formant energy modification showed intelligibility increases over conv/norm across talkers and listeners. For the male speaker, the combined modification of formant energies and fundamental frequency showed some improvement. Hearing impaired listeners did not increase their intelligibility scores for any modification across listeners and talkers.

While increasing energy around formant frequencies showed some promise, it was not investigated how it compares to simple processing by a carefully chosen LTI filter. This modification can be thought of as performing a kind of time-varying linear filter which is only non-unity during voiced portions of speech when it amplifies some portion of the spectrum (dependent on location of the formant frequencies). It may be possible that similar results could be achieved with a carefully chosen constant linear filter.

2.4 Summary

While some differences between clear speech and conversational speech have been identified, it is still unclear which directly lead to the intelligibility benefits of clear speech and how they can be introduced into conversational speech for a general enhancement system. Although speaking rate is drastically (and consistently) reduced

in clear speech, it has been proven that highly intelligible speech can occur at normal rates. Previous attempts at a general transformation scheme have highlighted several difficulties (such as the possibility that speakers use different strategies in eliciting clear speech) that must be addressed in future research.

Chapter 3

Investigation of Clear Speech Properties

3.1 Objectives

This thesis offers another analysis of clear and conversational speech to determine quantitative differences between the two modes and what possibilities may exist for utilizing them for speech enhancement. The approach is to look very closely at a few examples in the database and to attempt modifications of the conv/norm examples in those cases. If intelligibility enhancement "by-hand" in just a few cases is successful, it would be possible to attribute some amount of the benefits of clear speech to the particular characteristics altered in the transformation.

In previous studies, intelligibility test results were mainly used to identify average intelligibility levels for large classes of utterances (e.g. the overall intelligibility of clear/norm for a given speaker). Then various acoustical characteristics were measured and typically averaged across relevant occurrences in all utterances within a class. This thesis intends to take a closer look at the intelligibility test results to determine where errors occurred, and how the speech waveform differs at those locations.

The current objectives are best stated by posing several questions which the next two chapters will try to answer:

1. Which sentences provided a large increase in intelligibility?

2. In these successful sentences, what words or phrases provided this increase in intelligibility?

3. Can the intelligibility increase be attributed to even finer segments than words (e.g. phones)?

4. What quantitative features of highly intelligible words (or phones) differ from their less intelligible counterparts?

5. To what extent can signal processing be used to artificially reproduce these features in conversational speech?

These questions will be explored using two steps. First, an examination of the data and the test results will determine where the intelligibility improvements took place. After these "successful" portions are identified, properties of the speech signal during these segments will be analyzed for significant differences between conv/norm and clear/norm. These two steps are done at three levels of detail: sentence level, word level, and phoneme level. The results in this chapter will provide justification for the signal modification described in Chapter 4.

3.2 Database

The data used for this thesis was the clear/norm and conv/norm data collected by Krause [2]. It consists of two speakers: RG (female) and SA (male). RG and SA were chosen out of 15 originally recorded speakers based on their ability to effectively control their clarity and rate of speech. Each speaker read 50 pairs of sentences, resulting in 200 total utterances. Each sentence pair consists of a given sentence spoken in two modes: conv/norm (conversational at normal rates) and clear/norm (clear speech at normal rates). Each utterance was a grammatically correct, but semantically nonsense English sentence.

Each sentence contained either three or four keywords, which are defined as the nouns, adjectives, and adverbs in the sentence. For the intelligibility tests, all utterances were normalized to have the same long-term RMS value (excluding silences) and presented to eight normal hearing listeners in the presence of speech-shaped noise with SNR of -1.8 dB. Listeners were asked to write down the utterance as they heard it. The intelligibility of each sentence is defined as the percentage of keywords correctly identified, averaged across all listeners.

3.3 Sources of Intelligibility Increase

The first stage of analysis is concerned with investigating the results of the intelligibility tests performed on the clear/norm and conv/clear sentence pairs. Tests on this database have indicated that keyword identification in clear/norm was 14 points higher than in conv/norm when averaged across all keywords, all sentences, and both speakers [2]. This section deals with finding how this intelligibility increase is distributed amongst various sentences, words, and phones.

3.3.1 Sentence Level

Of the original set of 100 sentence pairs, it was desired to choose several to look at more closely. In selecting sentences, it would be appropriate to choose the "most successful" examples of clear/norm, i.e. those which substantially increased intelligibility over their conv/norm counterparts while maintaining a similar speaking rate. To determine this, the intelligibility increases and duration increases for each example of clear/norm were tabulated. The results are shown in Figure 3-1.

For each of the 100 utterance pairs, Figure 3-1 contains a point specifying the change in intelligibility and the percentage change in duration between conv/norm and clear/norm. The large circle and large X represent the averages for each speaker and the symbol of an X in a circle shows the average across both talkers and all sentences. Speaker SA had nearly double the improvement in intelligibility, but did so at a slight increase in sentence duration. On average, RG had almost no change

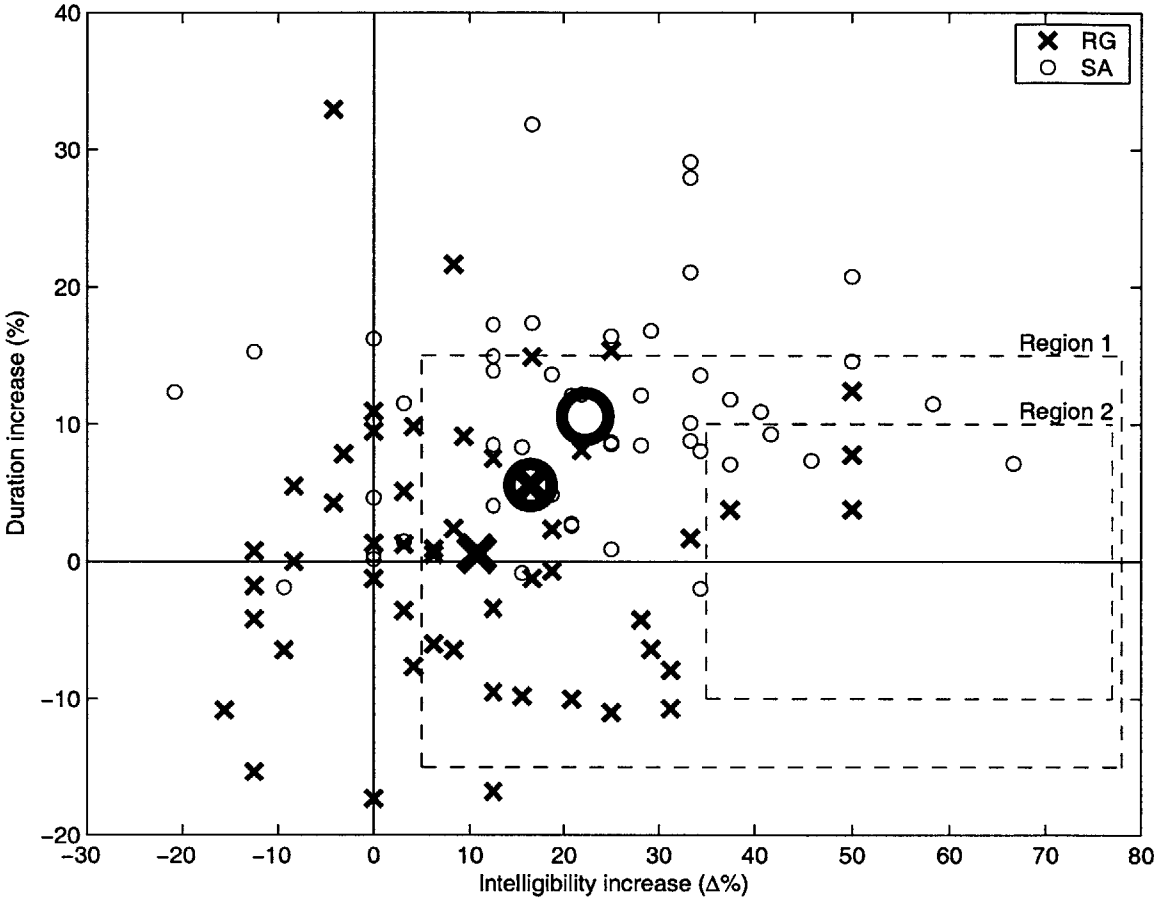

Figure 3-1: Performance of each sentence pair in clear/norm database. For each sentence pair, a point is plotted denoting the intelligibility increase and duration increase of clear/norm over conv/norm.

in duration, but had a smaller improvement in intelligibility of about 11 points.

The criteria used to define clear speech at normal rates used in past studies has been (i) an intelligibility increase of at least 5 points and (ii) a duration change of less then 15% [2]. The box labeled "Region 1" in Figure 3-1 encloses all sentence pairs meeting the criteria which contains 26/50 sentences for RG and 31/50 for SA. Therefore, only 57% of the sentences in the clear/norm database actually meet the formal standards for clear speech at normal rates. In fact, 25% of clear/norm examples have intelligibility scores less than or equal to those of conv/norm. Thus, using this entire database to determine the statistics of clear speech may "water down" the data with examples of unsuccessful clear speech, leading to difficulties in pinpointing important differences.

The group of sentence pairs chosen for close analysis and modification were those with the most substantial improvement in intelligibility without a significant change in duration. The box labeled "Region 2" in Figure 3-1 contains the seven sentences chosen. They are those with at least a 35 point improvement in intelligibility and a duration change of less than 10%.

3.3.2 Word Level

Given that a sentence has a particular intelligibility increase does not imply that some desired property is present throughout the sentence and the intelligibility improvement is uniform from beginning to end. Instead, the intelligibility increase may be due to the modification of a single keyword or phrase. Using reasoning similar to the sentence level selection, the analysis could focus on only those words which significantly contribute to the intelligibility improvements. This would prevent including words in the clear/norm database which are not actually more intelligible than their conversational counterparts.

Figure 3-2 is similar to Figure 3-1, but each point represents a keyword instead of a sentence. For the seven sentence pairs chosen to analyze, six of them had three keywords and one had four keywords, thus Figure 3-2 contains a total of 22 points. Again, "Region 1" specifies the boundaries for the definition of clear speech at nor-

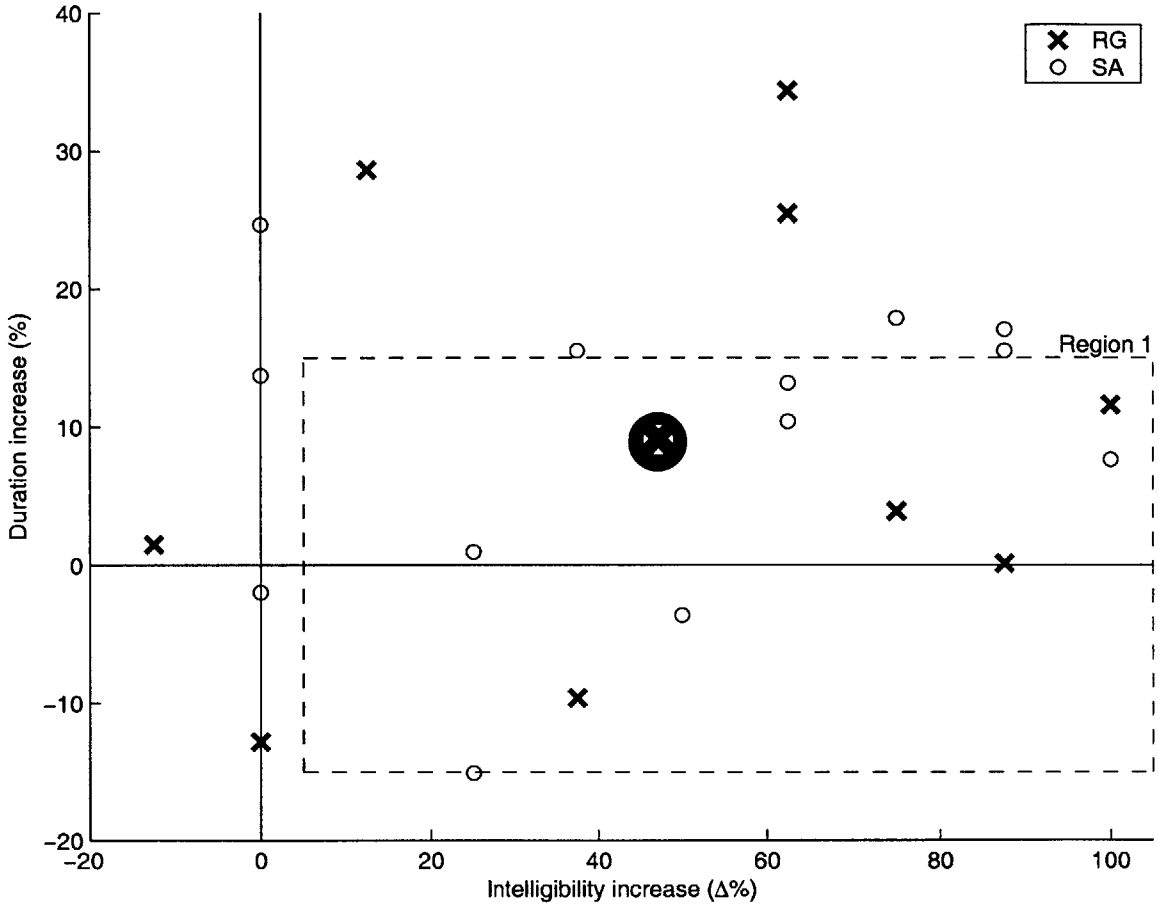

Figure 3-2: Performance of each keyword in the seven selected sentences. For each keyword, a point is plotted denoting the intelligibility increase and duration increase of clear/norm over conv/norm. The average for each speaker is shown by a large symbol (they are overlapping in this case).

mal rates. Despite using only keywords from the seven most successful examples of clear/norm, only 45% (10/22) of these words fall within the defined boundaries of clear speech at normal rates. Notice that words not contained in Region 1 are more likely to miss the criteria due to large changes in duration rather than insufficient intelligibility increases. This is in contrast to the sentence level, in which sentences were more likely to miss the criteria due to insufficient intelligibility increases.

Despite the large range in intelligibility increases, all 22 words will be kept for analysis. This is done because individual word scores are not necessarily independent of neighboring words within the sentence. Using nonsense sentences helped to minimize such effects, but because the sentences are syntactically correct and contain actual English words, some dependencies may exist.

While Figure 3-2 clearly shows which individual words are most successful, it is instructive to determine what properties these words have in common. If word intelligibility is highly correlated with some property, a generalized statement about which words receive the greatest improvement in clear/norm could be proposed. There are many ways to investigate such properties, several of which are explored here.

3.3.2.1 Effects of Sentence Position

One way to classify keywords is by their relative location in the sentence. Figure 3-3 plots the average intelligibility of keywords grouped by their relative location in the sentence[1]. Intelligibility is fairly constant in clear/norm, while for conv/norm the intelligibility drops drastically throughout the sentence. Despite the two speakers employing different strategies for eliciting clear speech, they have strikingly similar behavior in terms of how intelligibility varies throughout the sentence.

Figure 3-4 is produced by simply subtracting pairs of points in Figure 3-3 and shows the intelligibility increase as a function of sentence location. Figures 3-3 and 3-4 suggest that the speakers "trail off" in clarity in conversational speech as the sentence progresses. In clear speech, however, an approximately constant level of

[1]Throughout this thesis, keywords will be grouped by their relative sentence position: first, second, or third. For the single sentence containing four keywords, its fourth keyword is grouped in the third group, and its third keyword is ignored.

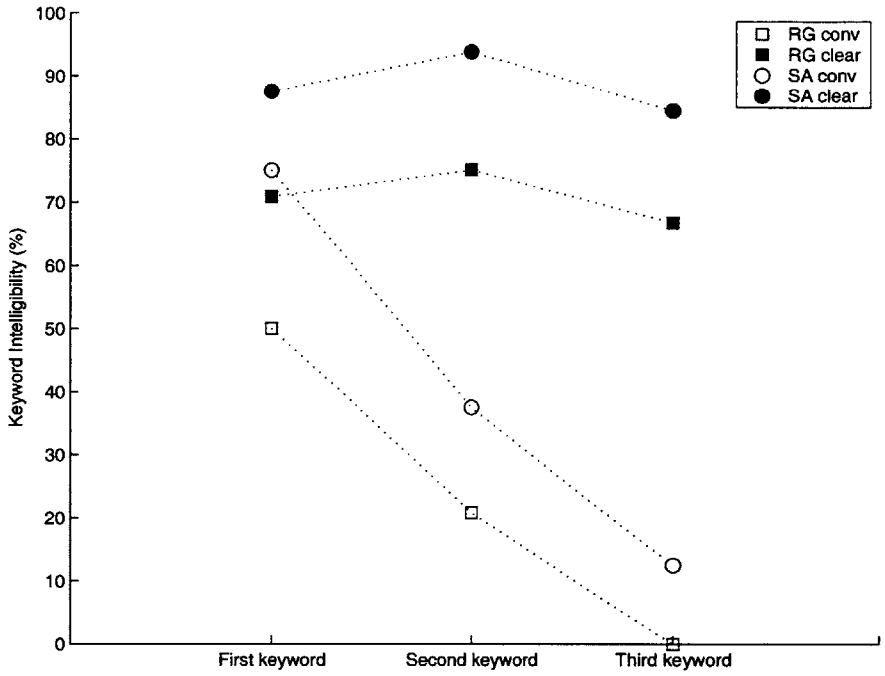

Figure 3-3: Intelligibility scores of individual keywords, grouped by their position within a sentence, speaker, and mode.

word clarity is sustained. Acoustical characteristics which may cause this effect will be explored later in the chapter.

Picheny *et al.* had reported a significant drop in intelligibility from the initial keyword to the following, but also noted that the improvement of the clear speech score over the conversational score was approximately the same for the two words [6]. Figure 3-4 shows that this is clearly not the case for this data. This discrepancy could either reflect a difference between clear/norm and clear/slow, or it could be a result of only analyzing a small number of the most successful sentences.

With intelligibility having such a strong dependence on word location for these seven sentences, it was decided to see how this trend applies to the other 93 sentence pairs in the database. Word scores were tabulated for all 200 utterances and placed into three groups according to the amount of intelligibility increase seen in clear/norm compared to conv/norm[2]. Results of this analysis are shown in Figure 3-5.

[2]Group 1 consisted of those with intell. increase of over 35 points and included 8 and 18 sentences for RG and SA, respectively. Group 2 had intell. increase between 5 and 35 points, with 50 and 62 sentences in RG and SA. Group 3 had intell. increases less than 5 points, with 42 and 20 sentences

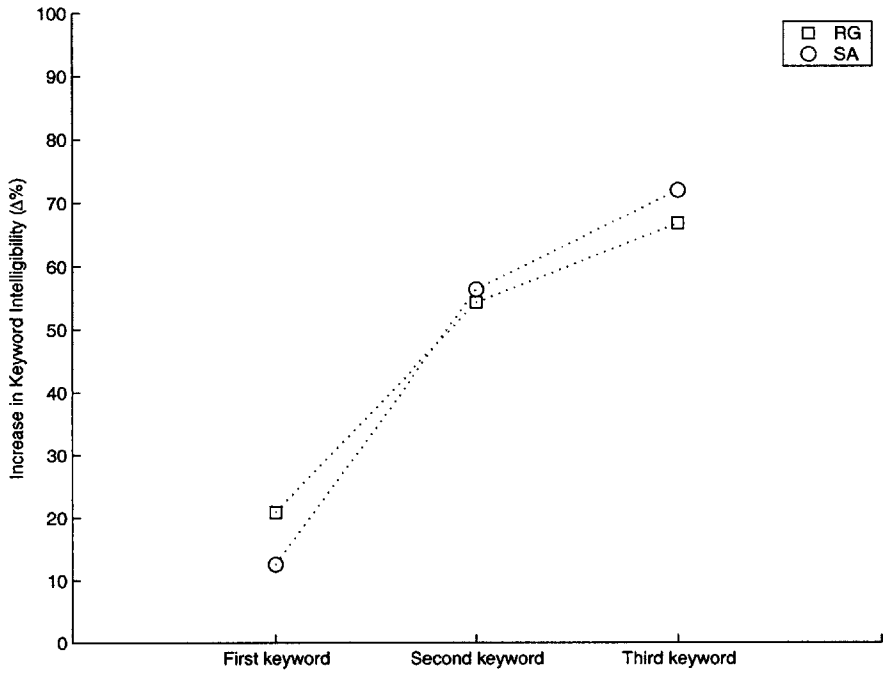

Figure 3-4: Intelligibility increase of clear/norm over conv/norm for individual keywords, grouped by sentence position.

Notice that nearly every line segment in Figure 3-5 has a negative slope. This means that across all classes, modes, and speakers, the intelligibility of a given keyword is almost always (on average) less intelligible than the one before it. Therefore, the "trailing-off" phenomenon is very widespread. However, in all three classes the clear/norm does not drop as much as conv/norm, which supports the theory that a major source of clear/norm's intelligibility benefit comes from minimizing this "trailing-off" effect. Figure 3-5 also reveals insight into how the low (or negative) intelligibility increases of some sentence pairs comes about. Sentences having lower intelligibility differences between clear/norm and conv/norm owe this to both decreased intelligibility of clear/norm and to increased intelligibility of conv/norm, rather than being dominated by one of these two factors.

from RG and SA

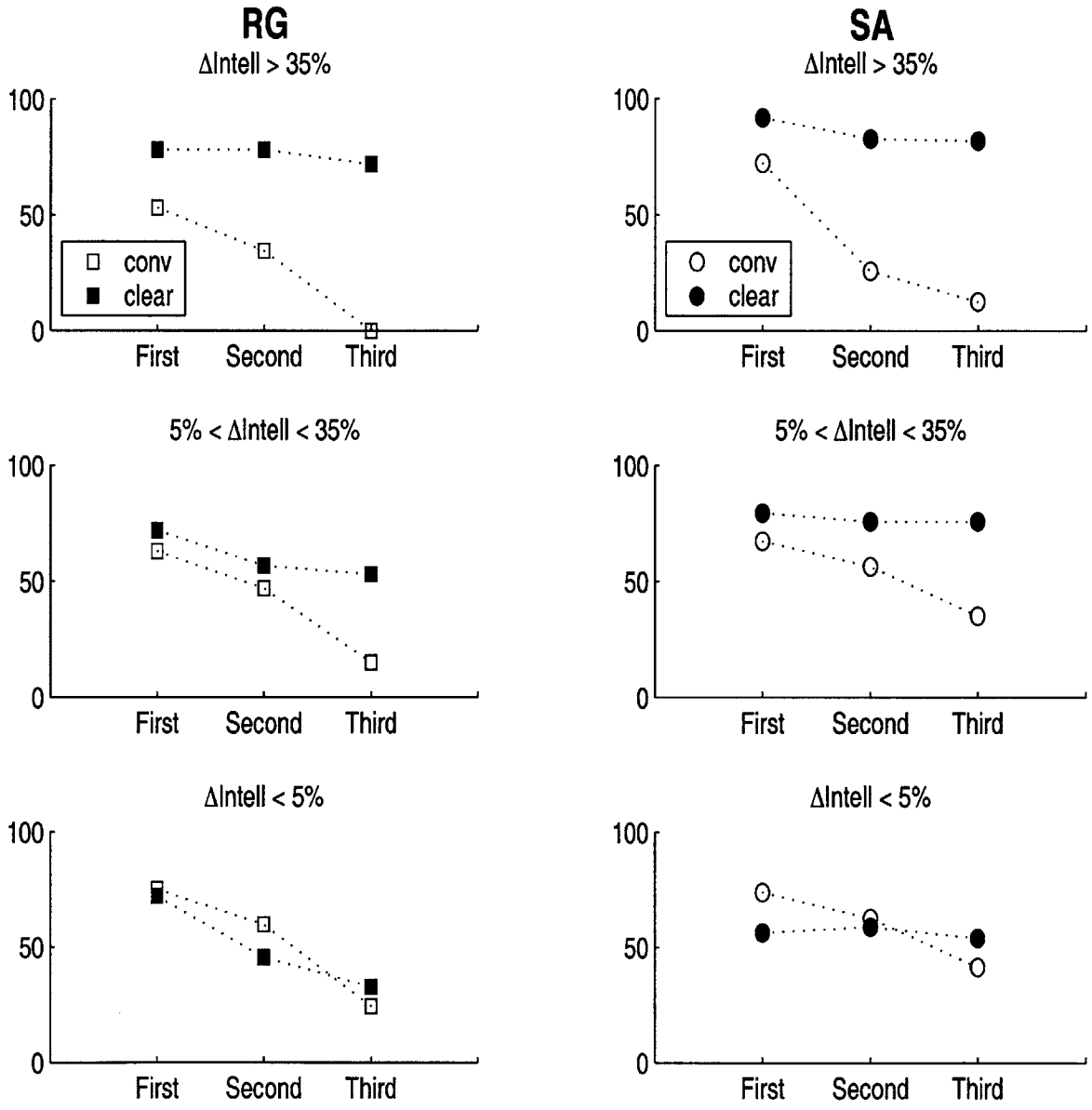

Figure 3-5: Keyword Intelligibility scores as a function of relative sentence position. Left column for speaker RG and right column for SA. Rows divide the database into 3 classes based on "success" of clear/norm, i.e. how large of an intelligibility increase clear/norm had over conv/norm.

3.3.2.2 Effects of Length

Another possible categorization which could distinguish between successful and unsuccessful words is word length. Figure 3-6 plots the average intelligibility increase for keywords as a function of the number of phones in the word (using only words from the seven chosen sentence pairs). There is a fairly constant increase in intelligibility improvement as word length (in number of phones) is increased. This suggests more complex words receive greater benefit from clear speech.

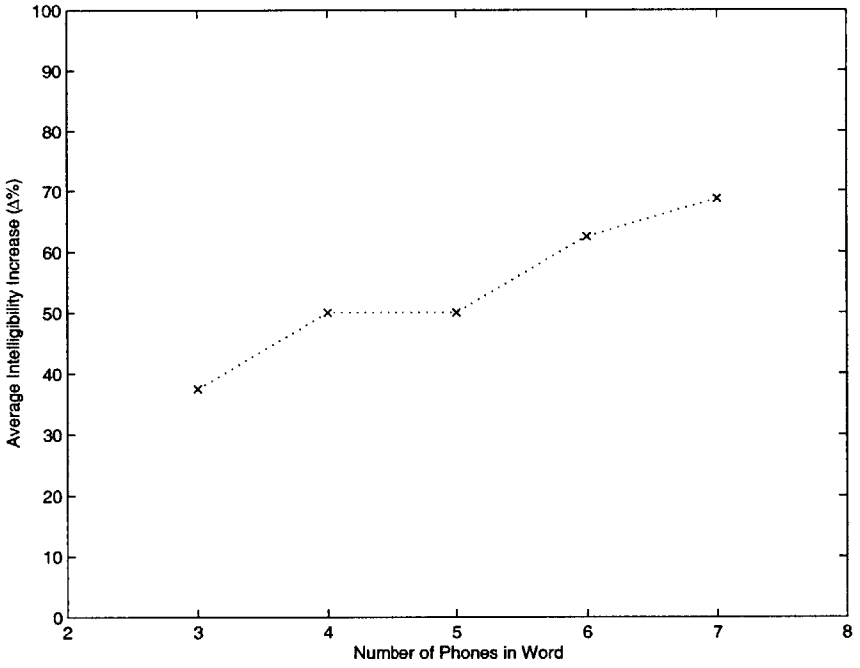

Figure 3-6: Keyword intelligibility improvement of clear/norm over conv/norm as a function of the number of phones in the word (for the seven selected sentence pairs).

3.3.2.3 Effects of Content

It is also reasonable to hypothesize that changes in word intelligibility could be influenced by the word's meaning. One well defined categorization of words that could be used is that of keywords versus non-keywords (i.e. nouns, adjectives, and adverbs vs. others). This, however, would be difficult to measure for several reasons. First, our very definition of intelligibility applies only to keywords. Also, non-keywords are

often interpreted by knowledge of neighboring keywords. Of course content can be classified in many other ways, but most meaningful classifications would be too difficult to quantify. However, there may be classifications which indeed could affect the amount of intelligibility increase a given word might receive. For example, a word's intelligibility improvement in clear speech may be dependent on how statistically common the words is.

3.3.3 Phoneme Level

The previous section demonstrated that for most sentences, the increased intelligibility of clear speech is far from uniformly distributed across all of its keywords. Likewise, the correct identification of a keyword can often be attributed to a small number of its constituent phones. Since a single phone insertion, deletion, or substitution labels the entire word incorrect, it is possible that a vast majority of phones analyzed in clear speech may have not been responsible for any intelligibility improvements.

One example is sentence RG-32, "Your rages could tell to a lobe." (keywords underlined). The word "rages" had intelligibility of 0% for conv/norm and 75% for clear/norm. (This is a rare case in which the first keyword provided a large intelligibility increase). However, with closer inspection of the test results, listeners always correctly identified the /r/ and the /e/ in conversational speech. The 0% correct identification was due entirely to the inability to identify the affricate consonant /jh/. Therefore, when looking for acoustical differences between clear and conversational speech that lead to the increased intelligibility of this example, we do not need to examine the /r/ and the /e/, but instead should concentrate on how characteristics of the /jh/ differed between conv/norm and clear/norm. This is just one example of many where most of a word's intelligibility increase was due to the identification of a single phone.

While specific cases like this can be examined, it is difficult to provide any generalized conclusions due to the large influence of context on phone identification within words. Intelligibility test results cannot be used to accurately monitor phone substitutions since subjects often interpolate individual phones to ensure that their response is

a valid English word. However, it appears that the aforementioned effect of "watering down" the data can also occur at the phoneme level, and it is likely very substantial when single phones are responsible for a large amount of intelligibility increase. In the RG-32 case, the /jh/ alone is responsible for approximately 50% of the intelligibility improvement of the sentence, but comprises only 7% of its phonetic makeup.

3.4 Causes of Intelligibility Increases

The previous section was concerned with determining what caused the increased intelligibility of clear/norm over conv/norm speech, i.e. which sentences, words, and phones were most successful in clear/norm. The next step is identifying why this occurred. Are there any general properties of the sentences, words, and phones which were found to be more successful in clear/norm? This section investigates this question at the same three levels of detail.

3.4.1 Sentence Level

Figure 3-1 identified which of the 100 sentence pairs exhibited a substantial difference in intelligibility between clear/norm and conv/norm. The next question becomes: What properties do the most successful sentences have in common? While this is a valid and interesting question, it will not be explored here. Analysis will be done on the seven chosen sentence pairs without a thorough investigation of what distinguishes these seven sentences from the other 93 *at the sentence level.* Exploring what makes some sentences lend themselves better to clear speech improvements could be an area for future research.

3.4.2 Word Level

It has been shown that keyword intelligibility tends to drop as the sentence progresses in conversational speech but stays fairly constant in clear speech. This section attempts to discover word level characteristics which could explain this phenomenon

using several simple measurements.

3.4.2.1 Average RMS Levels

The intelligibility test results being used came from presenting the sentences to normal hearing listeners in the presence of speech-shaped stationary noise. The clear/norm and conv/norm utterances were normalized to have the same long term RMS value (while excluding pauses in the speech). However, this long-term normalization leaves some freedom to the speaker on how to distribute energy throughout the sentence. One explanation of the trailing-off in intelligibility is that an associated trailing-off in loudness, or power, occurs as well.

Figure 3-7 and Figure 3-8 show RMS levels for individual keywords as a function of their location within the sentence. Sentence location is normalized to the range 0-1. Best linear fit lines are also shown. Points cluster fairly closely in conv/norm (Figure 3-7), and word RMS levels decrease fairly linearly. This decrease looks very similar to the linear decrease in keyword intelligibility for conv/norm shown in Figure 3-3. In Figure 3-8, points for clear/norm are not as tightly clustered, therefore there seems to be less of a general trend in RMS level for clear/norm keywords. The linear fit lines suggest that in both modes, keyword RMS decreases throughout the sentence. However, this decrease is less drastic in clear/norm (the slope of the linear fit lines decrease from conv/norm to clear/norm by 48% for RG and 66% for SA). If RMS differences between the two modes are computed and grouped by their relative keyword position, this trend becomes much more visible. This is shown in Figure 3-9, which plots the percentage change in word RMS level as a function of relative keyword position.

Figure 3-4 has shown that intelligibility increases rise with sentence location, and Figure 3-9 has shown a corresponding rise in relative RMS levels with sentence location. A natural conclusion would be that the increased RMS levels (and thus increased signal-to-noise ratios) are directly responsible for the intelligibility increase (or some portion of it). Figure 3-10 directly explores the relationship between keyword intelligibility changes and keyword RMS level changes. The points are somewhat sporadic.

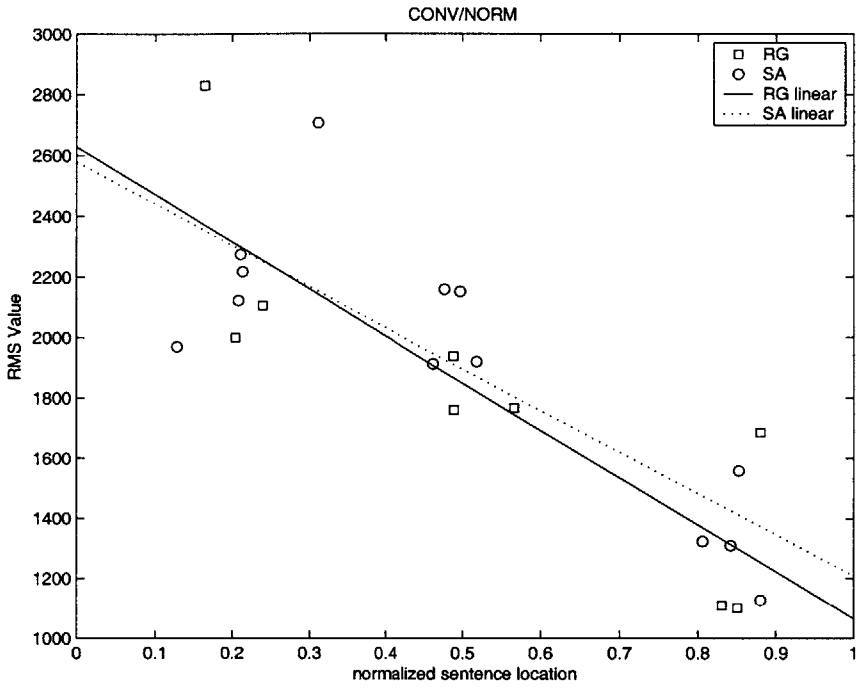

Figure 3-7: RMS levels of individual words in conv/norm as a function of their normalized position within a sentence

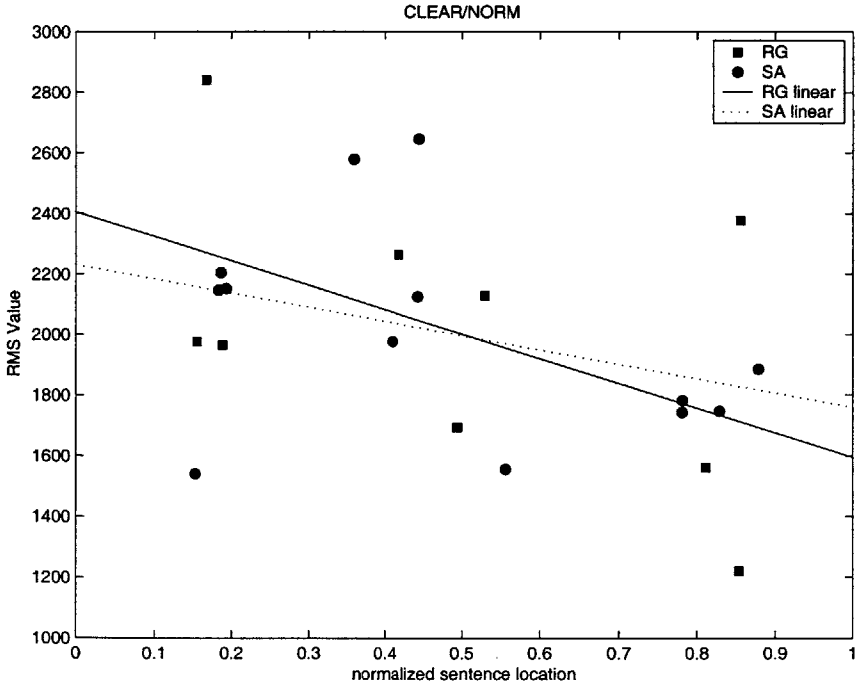

Figure 3-8: RMS levels of individual words in clear/norm as a function of their normalized position within a sentence

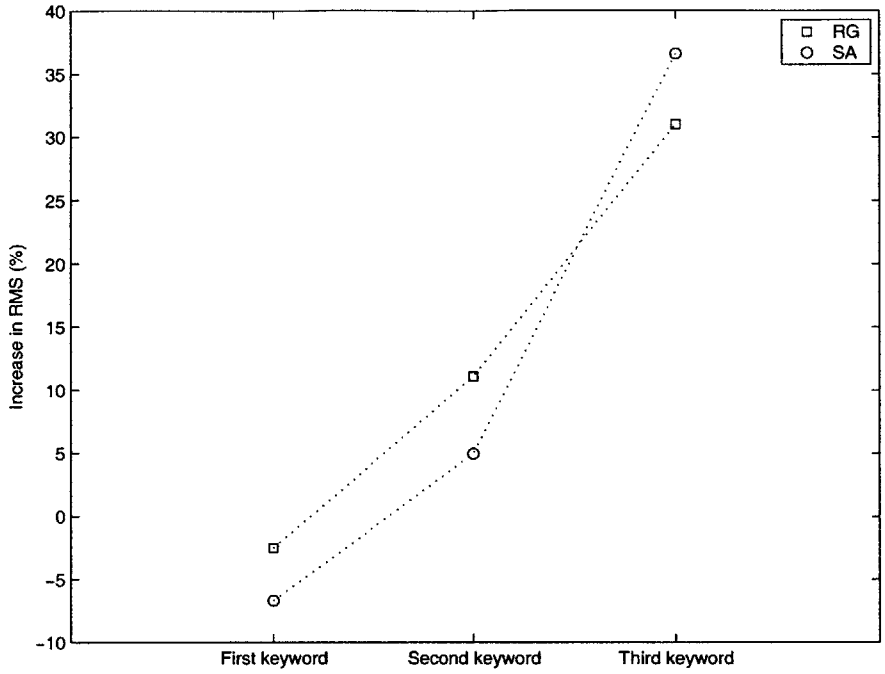

Figure 3-9: Percentage increase in keyword RMS level between clear/norm and conv/norm, grouped by relative sentence location.

However, the positive slopes of the linear fit lines do show increased intelligibility with increased RMS level, which would be expected.

There is also reason to believe that RMS level distributions may depend on content. While it is hard to quantify and compare which words are semantically more important, we can investigate RMS distributions between keywords and non-keywords. Figure 3-11 indicates how much (as a percentage) of each sentence's total energy is contained in the keywords. On average, this percentage increases from conv/norm to clear/norm for both speakers, indicating more energy is placed on keywords at the expense of non-keywords in clear speech. While this was implied by some of the previous plots in the chapter, this directly shows that the redistribution of energy in clear speech has a dependence on the word meaning.

3.4.2.2 Durations and speaking rates

While the successful clear/norm utterances were at approximately the same speaking rate as conv/norm, there could be significant variation on word durations within

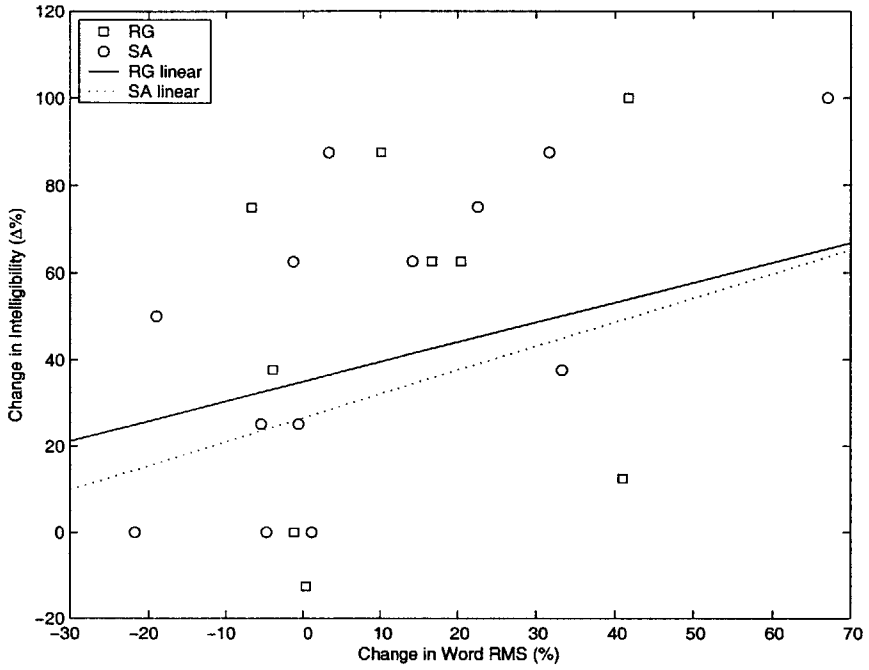

Figure 3-10: Keyword intelligibility increase as a function of keyword RMS level increase between clear/norm and conv/norm.

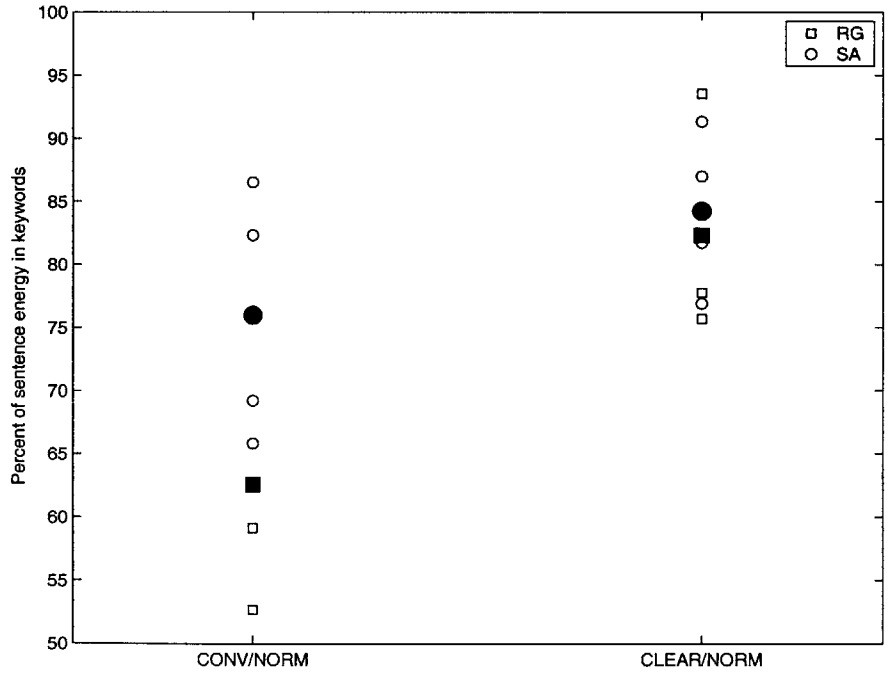

Figure 3-11: For each utterance, a point is plotted indicating what percentage of the utterance's total energy is contained in its keywords. Shaded markers indicate average values for each speaker in each mode.

each sentence. With reasoning parallel to that of the word RMS levels, durations of some words may be lengthened at the expense of others, while keeping long-term (sentence-level) speaking rates approximately constant.

Figure 3-12 shows the percentage change in keyword duration vs. relative position in the sentence. For the seven sentences used to generate this plot, the sentence level durations of clear speech were 8% longer, on average, than conversational. The second and third keywords are lengthened relative to the long-term speaking rates. This is evidence that the reduction in "trailing-off" in clarity done in clear speech is a result of lengthening words as well as increasing intensity. Once again, this trend shows a strong similarity between the two speakers.

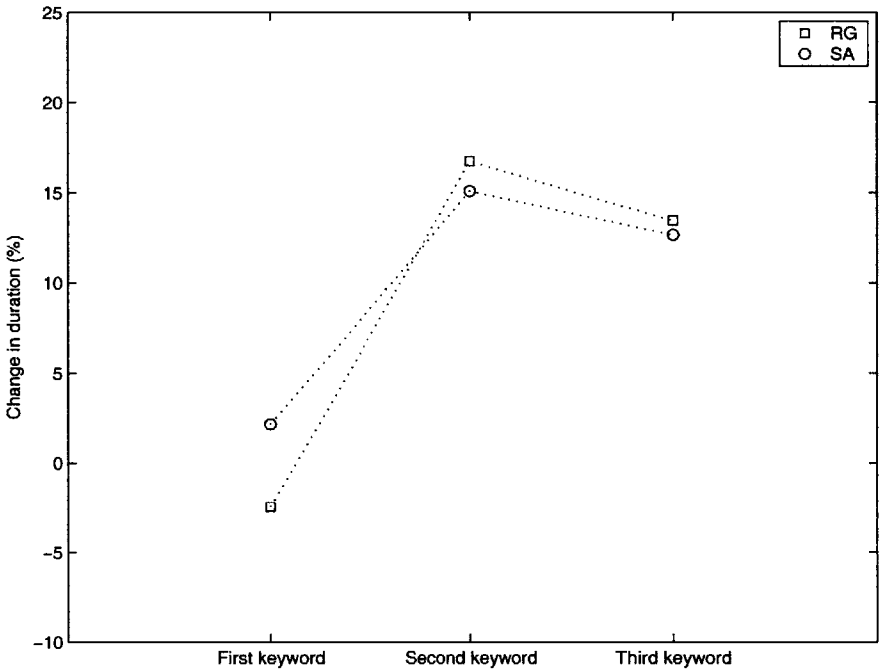

Figure 3-12: Percentage change in keyword duration between clear/norm and conv/norm, grouped by relative sentence location.

The experiments of Picheny *et al.* had shown that speaking rate alone does not seem to contribute to the high intelligibility of clear speech [12]. So, while lengthening words may not directly increase intelligibility, some associated acoustic properties of lengthening may contribute. Also, such effects may be part of a fundamental difference between clear speech at normal rates and clear speech at reduced rates.

The effect of content on durations was also investigated. Unlike power levels, it appears that keyword durations are not increased relative to non-keywords in clear speech. This is illustrated in Figure 3-13 below. This indicates that the lengthening of the second and third keywords comes at an expense of the first keyword and a slight overall duration increase, rather than at the expense of non-keywords.

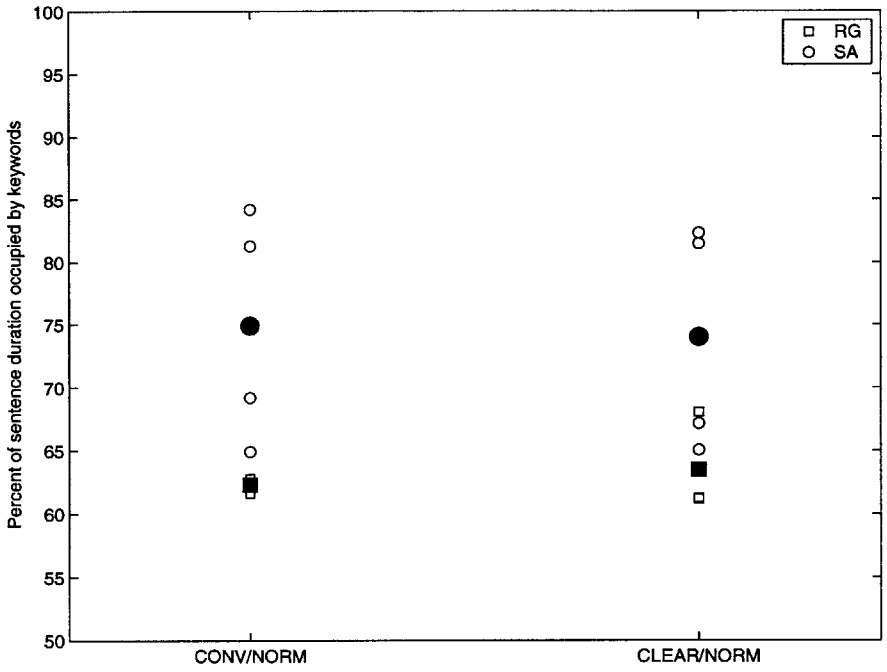

Figure 3-13: For each utterance, a point is plotted indicating what percentage of the utterance's total duration is occupied by its keywords. Shaded markers indicate average values for each speaker in each mode.

3.4.2.3 Distribution of Energy

Finally, it is worth briefly looking at total energies at the word level. Figures 3-14 and 3-15 relate keyword energy to intelligibility and sentence position. These results are not surprising considering the results from power and duration measurements, but the amount of change shown in Figure 3-15 is quite dramatic. On average, the final word in the sentence has about 100% more energy in clear/norm than in conv/norm, for both speakers.

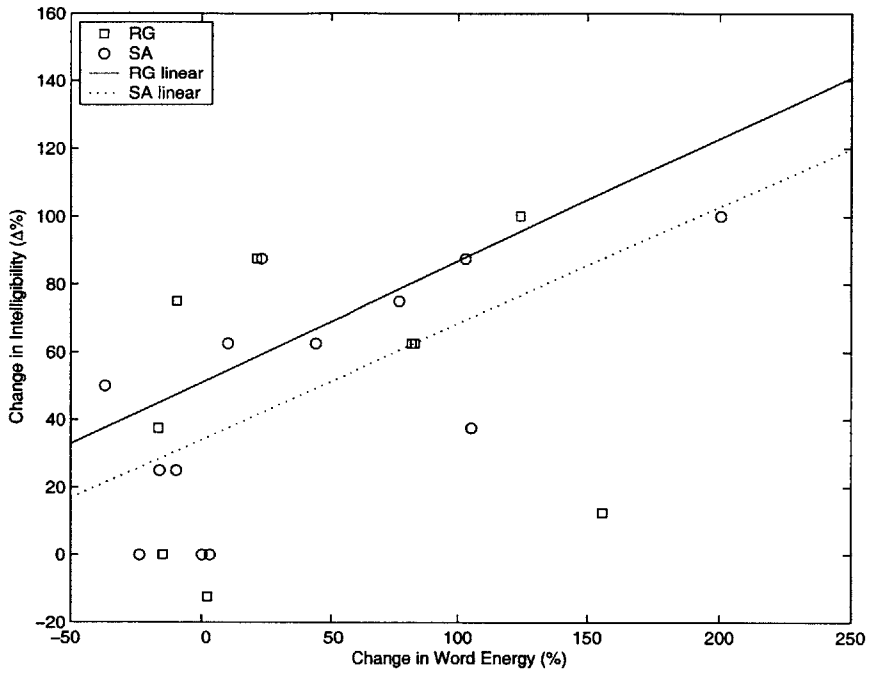

Figure 3-14: Keyword intelligibility increase as a function of keyword's total energy increase between clear/norm and conv/norm.

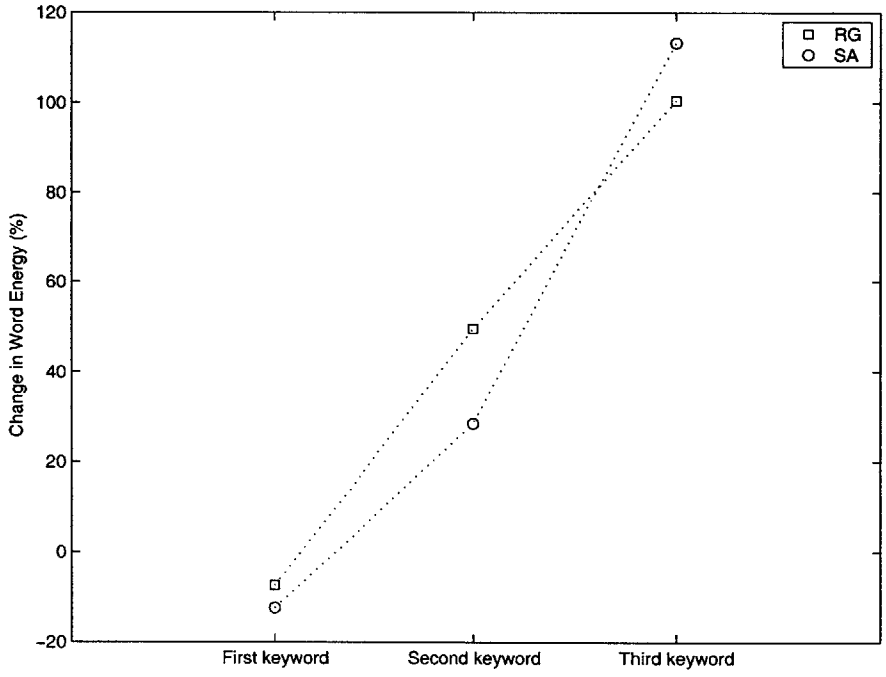

Figure 3-15: Percentage increase in keyword energy between clear/norm and conv/norm, grouped by relative sentence location.

3.4.3 Phoneme Level

Because it is hard to identify intelligibility changes for particular phones, common acoustic properties of highly intelligible phones cannot be explored. However, it is possible to look at several properties at the level of phones, including energy, duration, and fundamental frequency.

3.4.3.1 Distributions of Energy and Duration

Extensive analysis of energies and durations of phones and various phone classes has already been performed on the clear/norm database [3]. While no significant differences were found between conv/norm and clear/norm, it is possible that changes in these quantities depend more on the phone's location than on its type. One way to measure such dependencies is to look at segmental power, energy, and duration as a function of that segment's intra-word position. Figure 3-16 shows percent changes in these quantities between clear/norm and conv/norm, grouped into four intra-word locations (from the beginning of the word at left, to the end of the word at right). For each of the four groupings, all segments whose midpoint was within that range were grouped together and their properties were averaged. This figure serves as a rough contour of how changes in these quantities behave as a function of their location in a word. This reveals that word-final and word-initial segments receive a considerable boost in amplitude. It is likely that the large increase at the end of the word (word-final phones have an average change in energy of 350%) is due largely to the much louder release of word-final stop consonants.

3.4.3.2 Fundamental Frequency

Fundamental frequency was one of the three factors Krause decided to include as a possible cause for the high intelligibility of clear speech [3]. While experiments with pitch-modified conv/norm have shown somewhat limited improvements on intelligibility, it is constructive to analyze the pitch behavior of the seven most successful utterance pairs.

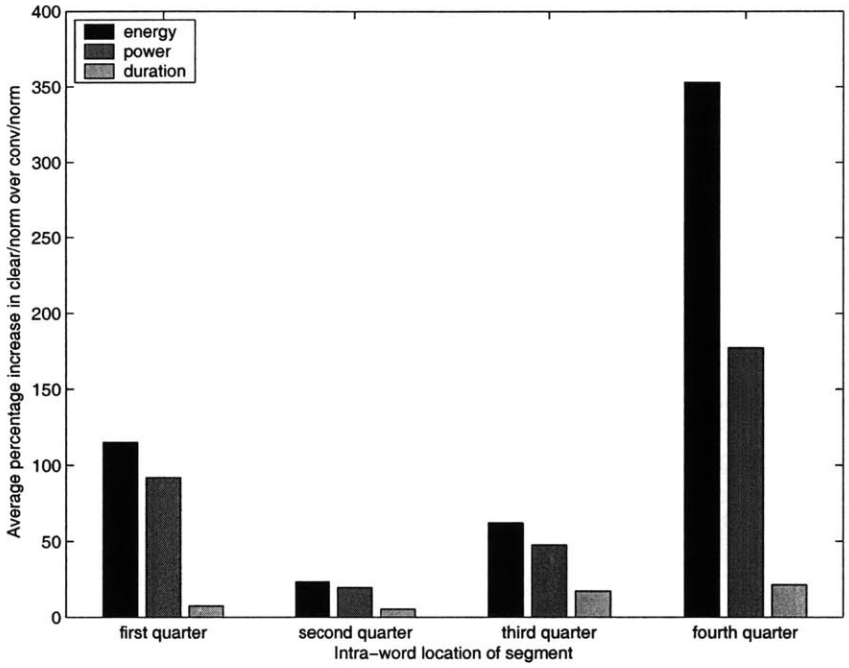

Figure 3-16: Differences in segment level energy, duration, and power between clear/norm and conv/norm as a function of the segment's intra-word location. Keyword segments were grouped into one of four positions based on the segment's position within the keyword.

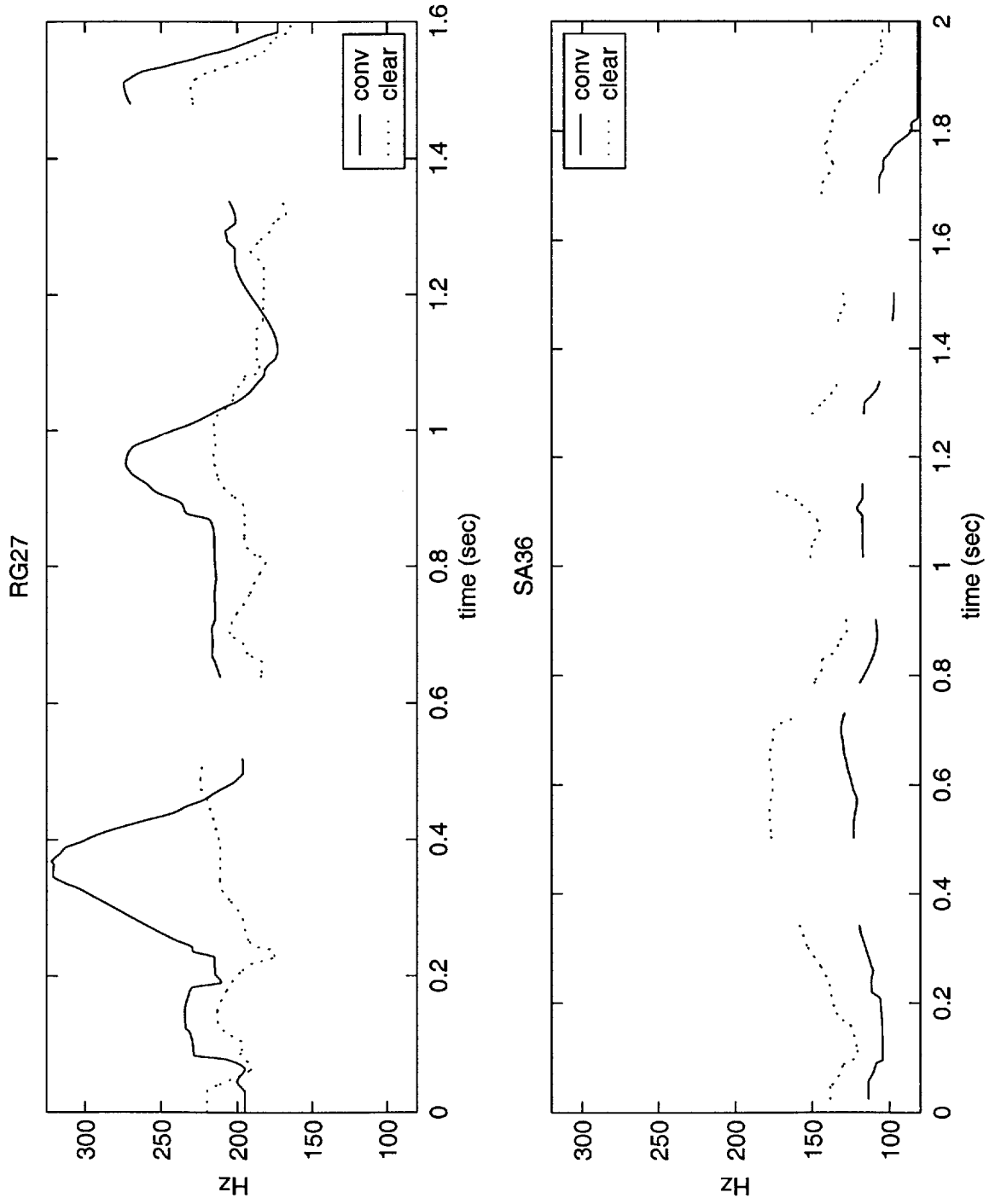

Figure 3-17: Example pitch contours for each speaker and both modes.

Figure 3-17 shows an example pitch contour for each speaker. The pitch contour for clear/norm is scaled so that segment boundaries are time-aligned with conv/norm for comparison. Pitch differences between clear/norm and conv/norm are very speaker-dependent. The most noticeable and consistent trend is the increase in pitch for speaker SA. Clear speech for speaker RG had a slightly lower pitch, as well as a decrease in large fluctuations. Table 3.1 shows the means and standard deviations for the seven sentences analyzed.

Table 3.1: Means and standard deviations of fundamental frequency measurements for the seven selected utterance pairs. Relative changes between clear/norm and conv/norm given in parentheses.

	RG		SA	
	conv/norm	clear/norm	conv/norm	clear/norm
Mean	216	199 (-8%)	105	140 (+33%)
Std	32.2	19.7 (-39%)	13.7	19.9 (+45%)

Subjective listening of the utterances led to the conclusion that both speakers use pitch mainly to alter the tone of their speech. RG has large pitch variations in conv/norm leading to a very "light and friendly" tone. In her clear speech, RG employs a more direct and steady pitch, resulting in a more serious tone. On the other hand, SA greatly increases pitch in clear/norm suggesting increased vocal strain. These subjective features could be side-effects of a particular tone that the speakers associate with situations requiring the production of clear speech. Also, with SA consistently raising pitch and RG lowering pitch, it may suggest that both are approaching some target pitch for optimal intelligibility in the range of 150-200 Hz.

3.5 Summary

The first part of this chapter served two purposes: (i) to select a small number of sentences to attempt transformations on and (ii) to attempt to better understand where intelligibility increases take place. Investigation of the test results has revealed that much of the success of clear speech may be attributed to a small fraction of the

clear/norm utterance. Therefore, previous studies may have diminished the evidence of acoustical differences between clear and conversational speech by including elements which offered no increase (or in some cases a decrease) in intelligibility. It was also shown that keyword intelligibility improvements have a strong dependence on their relative location within a sentence.

The second half of the chapter tried to explain these results and led to the hypothesis that much of the intelligibility advantages of clear/norm are due to two general factors: (i) speakers do not "trail-off" in loudness and enunciation as the sentence progresses in clear speech, and (ii) speakers place stronger emphasis on "important" words (i.e. keywords) in clear speech. This emphasis may be realized as various combinations of changes in energy, duration, and possibly pitch.

Increasing power levels during keywords has previously been reported by Krause. The low frequency temporal envelope modulations which she included as a contributing factor of clear speech could be explained by increased energy in keywords, since keywords occur at a rate of approximately 3 Hz. Krause explicitly noted that "relative intensity of content words compared to function words was increased in clear/norm" and also noted that these increases were more pronounced at the ends of sentences [3]. However, it not surprising that an artificial boost of these modulations will not increase intelligibility because in general, the amplification will not line up accurately with the keywords to boost the correct part of the waveform. To test energy levels more precisely, the clear/norm waveform can be used to tailor amplifications specifically for the conv/norm utterance. Such a transformation is tested in the next chapter.

Chapter 4

Modification of Conversational Speech

The previous chapter suggested that clear speech owes its high intelligibility to the speaker placing more emphasis on words crucial to understanding (i.e. keywords) and sustaining enunciation and power levels throughout the course of the sentence. While some of this emphasis is likely due to various phonological phenomena, it is possible that a substantial amount is due simply to changes in RMS level. This chapter tests this hypothesis by redistributing the energy in conv/norm so that segmental power levels match those of clear/norm. Measuring the resulting changes in intelligibility will offer insight into how short-term RMS levels contribute to the benefits of clear speech at normal rates.

4.1 Modification

4.1.1 Segment Normalization

The seven chosen sentences were segmented by hand. Segmentation was not done into formal phonetic units, but somewhat arbitrarily into the smallest segments with some determinable boundary in both conv/norm and clear/norm by visual observation of the waveform and the spectrogram. On average, the sentences contained 27 segments.

Each segment in conv/norm was multiplied by a different constant so that the RMS level of each segment equals that of the corresponding segment in clear/norm. Occasionally, a segment occurred in the conv/norm utterance but not in the clear/norm utterance. These segments were left unchanged. Also, nothing was done for any segments appearing in clear/norm but not in conv/norm (including pauses). These steps insure that amplitude scaling is the only difference between the original and modified conv/norm. After normalizing each segment, the entire utterance was normalized so that its sentence level RMS value is equal to that of the original conv/norm utterance.

4.1.2 Other Modifications

Two other modifications were also tested in an attempt force conv/norm to resemble clear/norm. A sinusoidal analysis-synthesis system [9, 11] was used to alter the time-scale and fundamental frequency properties of conv/norm to match those of clear/norm. Segments in conv/norm were uniformly time-scaled to match the duration of the corresponding clear/norm segments, and sentence level pitch contours in conv/norm were replaced by those of clear/norm. While these more invasive modifications transformed conv/norm to sound similar to clear/norm, there were often unnatural side effects that could negate any intelligibility improvements which these procedures attempted to produce.

After some subjective testing on the effects of these modifications, it was decided to use only segment normalization for further study. This was done for two reasons. Most importantly, simple amplification caused no unwanted artifacts. For some examples, there was a noticeable degradation in quality or naturalness after modifying time-scale or pitch. Secondly, testing more than one modification could lead to learning effects in the test subjects since the same sentence must be presented multiple times. Therefore, testing only segment normalization offered the best opportunity to test a single characteristic (segmental power) without causing any side effects of signal processing, unnatural results, or learning effects.

4.2 Results

4.2.1 Preliminary Results

Because the chosen modification only involved simple amplification, there were no problems with the quality of the modified speech. However, due to the required long-term normalization and occasional segment miss-matching, the performed segment normalization did not ultimately leave every segment with RMS level equal to that of clear/norm. Figure 4-1 and Figure 4-2 show histograms for segmental energy and power differences between conv/norm and clear/norm before and after segment normalization is applied.

Ideally, the bottom histogram in Figure 4-1 would consist of a single bar above zero, which would indicate that after modification, the RMS levels of all segments are equal to those of the corresponding segments in clear/norm. However, about 46% of segments still have a slightly higher segmental RMS level in clear/norm, as seen in the "tail" on the positive side of this figure. While very few segments have a substantial difference in RMS level, one of the conclusions of Chapter 3 was that changes in a small number of segments can greatly affect the overall intelligibility.

In Figure 4-2, the distribution of segment energy differences is not changed as much by the modification since a given segment generally has different lengths in clear/norm and conv/norm. However, since extreme deviations in energy were due more to amplitude differences than to duration differences, outlying segments in Figure 4-2 are removed by the modification.

It was concluded with subjective listening that the modified conv/norm did seem to have a positive effect on intelligibility when presented in noise, although the extent was highly dependent on the particular sentence. In some cases, changes were virtually unnoticeable, while in others there was a clear improvement. Not surprisingly, the most noticeable improvements occurred when the final keyword in the sentence was amplified and thus was more easily heard above the background noise.

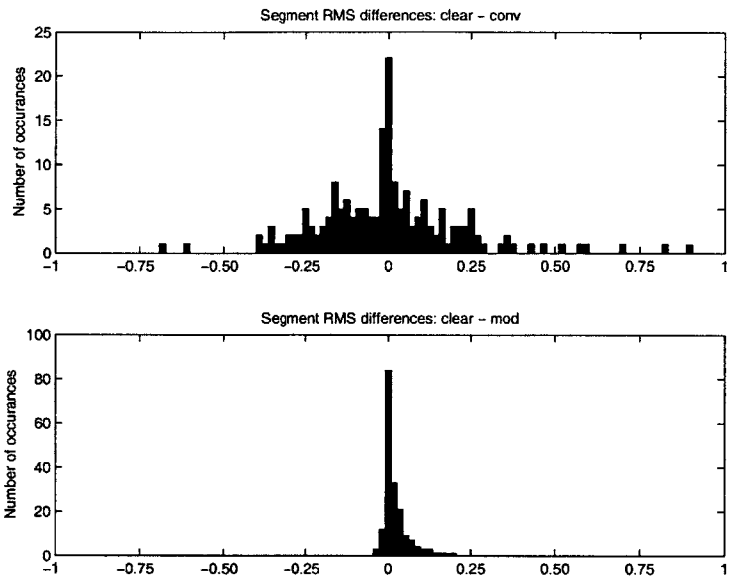

Figure 4-1: Histogram of segment level RMS differences. Top shows differences before processing (i.e. differences between clear/norm and conv/norm). Bottom shows differences after processing (i.e. differences between clear/norm and modified conv/norm). Units on the x-axis are fairly arbitrary and are thus shown on the range [-1,1].

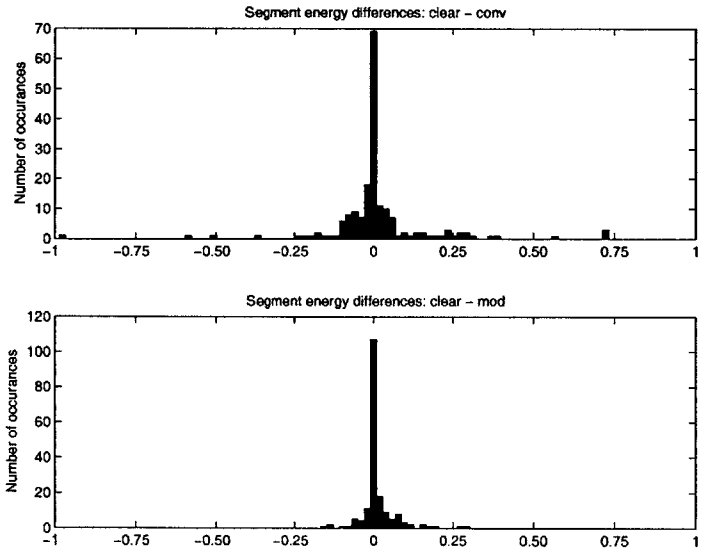

Figure 4-2: Histogram of segment level energy differences. Same layout as Figure 4-1, using energy instead of RMS level. Units on the x-axis are fairly arbitrary and are thus shown on the range [-1,1].

4.2.2 Intelligibility Tests

Five normal hearing listeners were presented with the original and the modified conv/norm utterances for each of the seven sentences in the presence of noise with SNR of -1.8 dB. This was done in two sessions separated by about two weeks. In an attempt to average out any learning effects or session inconsistencies, the original and the modified utterances were split between the two sessions. Intelligibility scores were defined as in previous experiments [6].

The original conv/norm utterances averaged 47% correct identification of keywords, and the segment normalized versions had a score of 55%. This 8 point increase is substantially lower than the increase between conv/norm and clear/norm measured in previous testing for these examples. However, this previous test [2] had identified the intelligibility of conv/norm to be 33% (for these seven sentences), which is considerably lower than the 47% measured here. Therefore, it was decided to retest the five subjects for clear/norm so that a fair comparison between all three modes could be performed. A third testing session presented all seven clear/norm utterances in the same conditions as before. The average intelligibility for this set was measured to be 77%.

A summary of testing results is shown in Figure 4-3. This data suggests that energy redistribution accounts for 8 points out of the 30 point improvement of clear/norm over conv/norm (8/30 = 26.7%). However, it is possible that the 77% score for clear/norm may be slightly high due to learning effects. By testing such a small number of sentences and presenting them three times, some familiarity of the sentences could have been acquired. The first two testing sessions consisted of a mix between conv/norm and modified conv/norm in an attempt to average out any learning effects. The third session consisted of all sentences in clear/norm. Therefore, this 77% score may be thought of as an upper bound on an intelligibility score for clear/norm.

Figure 4-4 shows intelligibility scores for the five individual listeners. Four of them show increased intelligibility for the segment normalized conv/norm; however, the amount of this increase varies significantly. The modified speech has a higher variance

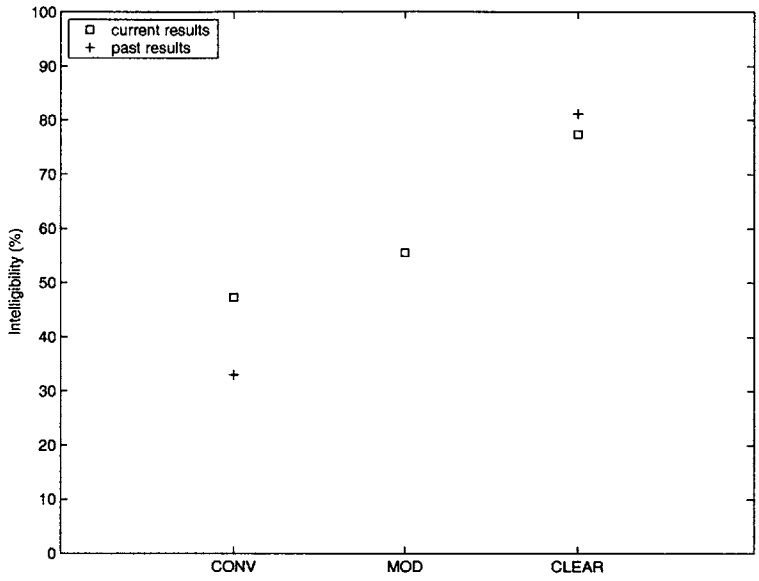

Figure 4-3: Intelligibility test results for the three modes tested. Past results indicate scores from [2] on the same set of sentences, under the same conditions.

in intelligibility scores between listeners than either conv/norm or clear/norm. This may imply that the effects of energy redistribution on intelligibility can significantly vary depending on the hearing ability of the listener.

It is useful to again look at keyword performance as a function of relative sentence location. Figure 4-5 plots intelligibility of the three modes grouped into three classes: first, second, or third keyword. The redistribution of energy transforms the shape of the plot for conv/norm to closely resemble that of clear/norm. Therefore, segmental power levels seem to account for the time-dependent differences in intelligibility between conv/norm and clear/norm. Other characteristics present throughout the sentence must be responsible for the remaining gap between modified conv/norm and clear/norm seen in Figure 4-5.

Finally, Figure 4-6 displays results for individual sentences, averaged across all listeners. For each sentence, the three bars represent an intelligibility improvement over conv/norm for each of three cases. The first bar shows the improvement resulting from the performed energy redistribution. The second and third bars represent the improvement seen by clear/norm over conv/norm from two different intelligibility

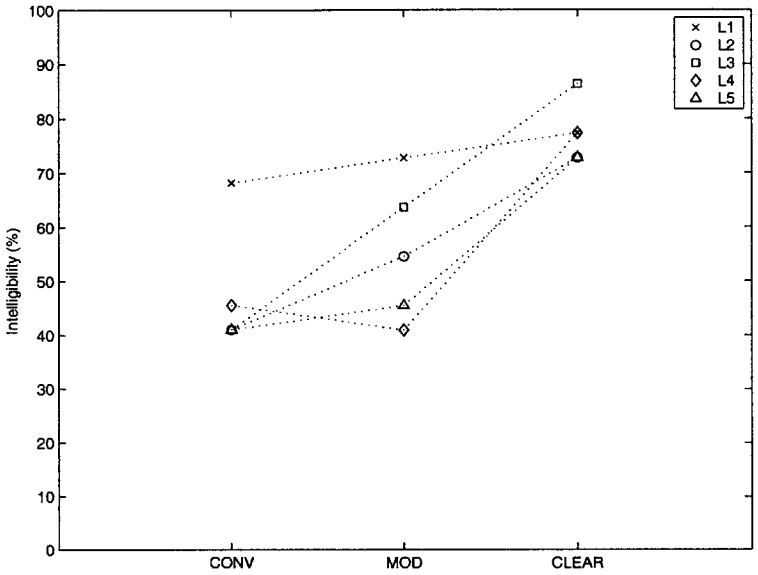

Figure 4-4: Intelligibility test results for individual listeners.

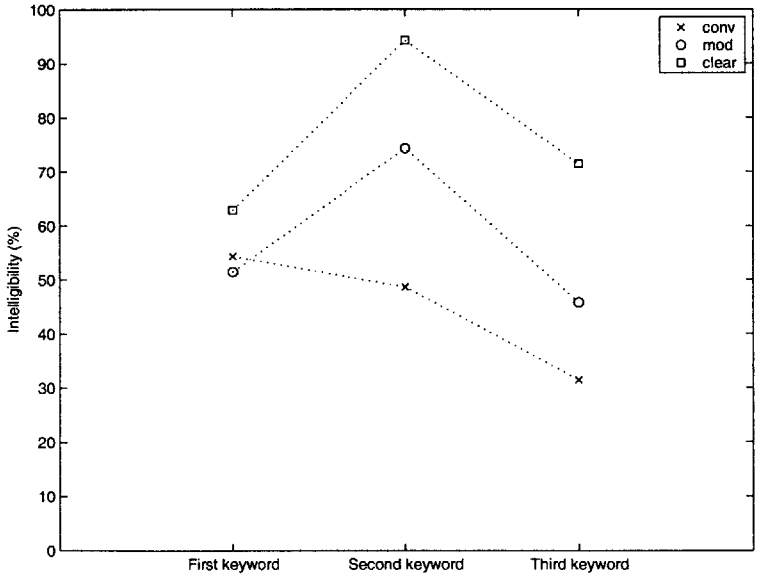

Figure 4-5: Intelligibility test results, grouped by relative location of keywords. Scores are averaged across both speakers and all five listeners.

tests (the current tests and those in [2]). Significant differences between these two measurements suggest that more thorough testing may be needed. Overall, the varying behavior of the seven examples warrants a more detailed analysis of how these intelligibility scores came about.

4.2.3 Detailed Results

Sentences RG-27, RG-32, and SA-19 exhibit behavior which is not surprising. The segment normalization results in increased intelligibility, but not to the extent that its intelligibility reaches the level of naturally produced clear/norm. This would indicate that power levels play a large role in clear speech, but there are other contributing factors. The other four examples exhibit somewhat unexpected behavior and are therefore looked at more closely in the following sections. To aid in the study of specific sentences, detailed results from the current intelligibility tests were tabulated and are presented in Table 4.1.

4.2.3.1 RG-37

RG-37 was investigated because the modified utterance resulted in an intelligibility score lower than that of the original conv/norm. This decrease in intelligibility is small enough to be attributed to inaccuracies in testing. (The 7 point decrease is equivalent to a single listener missing a single keyword). Therefore, it can be concluded that segment normalization had little effect on intelligibility despite clear/norm showing a 27 point increase. Table 4.1 reveals that this increase in clear/norm can be attributed to the 100% identification of the final keyword, "ramp" (as opposed to 20% in conv/norm). Therefore, this is a case where the trailing off in intelligibility at the end of a sentence cannot be partially attributed to a trailing off in RMS level. Subjective listening suggests that the high intelligibility of "ramp" in clear/norm may be due to greater retroflexion of the initial /r/ and a more distinct separation between the /r/ and the end of the previous word.

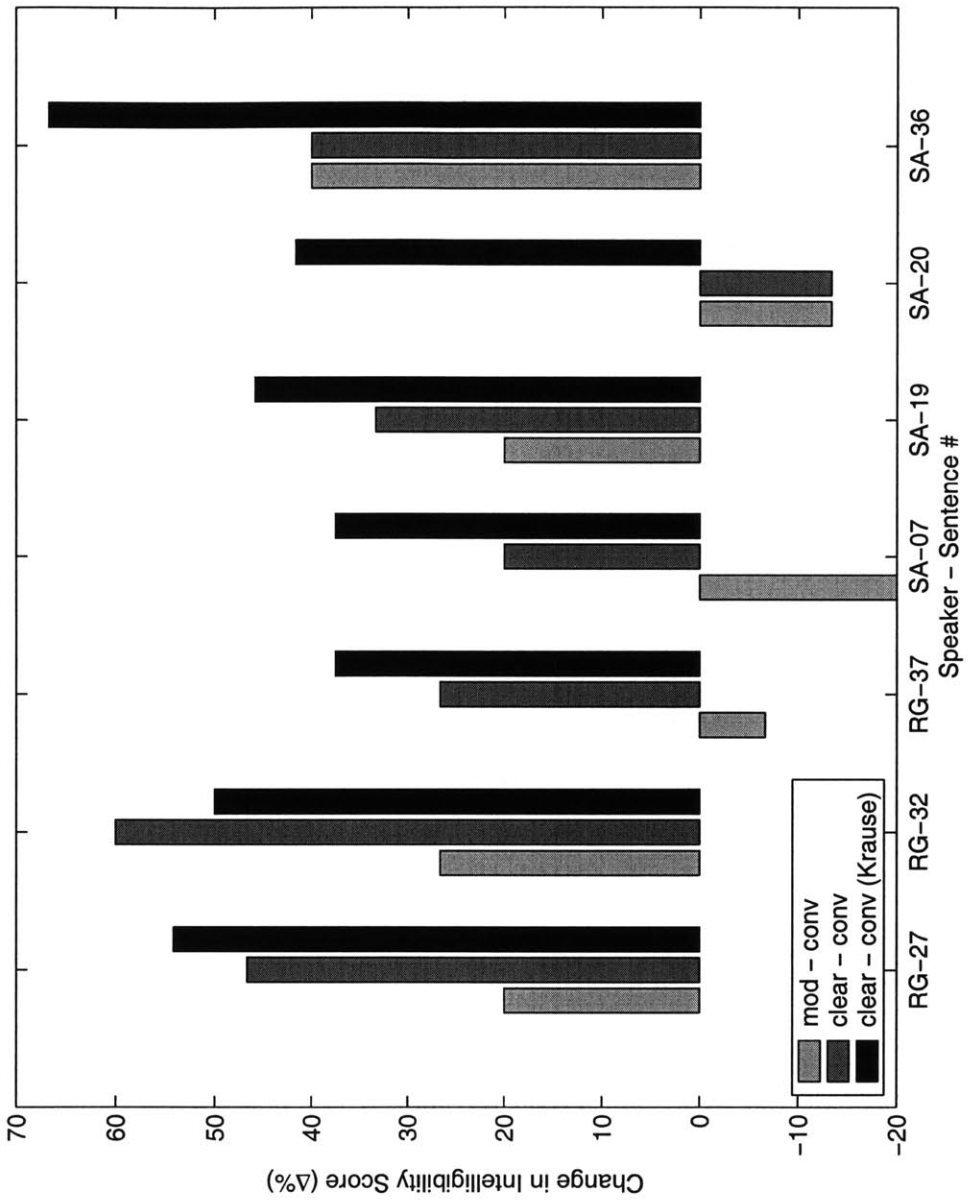

Figure 4-6: Intelligibility scores for the seven individual sentences studied. Each bar represents the amount of intelligibility improvement over conv/norm. The first bar in each group is the score from modified conv/norm minus score from conv/norm. Second bar the is score from clear/norm minus score from conv/norm. The third bar is identical to the second, but using results from [2].

Table 4.1: Detailed intelligibility scores (percentages of keywords identified correctly) for the example sentences averaged over the five listeners.

sentence	**mode**	**keywords**				**Ave**
RG-27		*back*	*kneel*	*quest*		
	conv	60	0	80		47
	mod	40	60	100		67
	clear	100	80	100		93
RG-32		*rages*	*tell*	*lobe*		
	conv	0	0	0		0
	mod	0	40	40		27
	clear	60	100	20		60
RG-37		*glow*	*soak*	*ramp*		
	conv	40	60	20		40
	mod	20	80	0		33
	clear	0	100	100		67
SA-07		*fierce*	*arrow*	*balances*	*debt*	
	conv	80	100	100	0	70
	mod	40	100	20	40	50
	clear	60	100	100	100	90
SA-19		*store*	*fall*	*breeze*		
	conv	80	40	0		40
	mod	100	60	20		60
	clear	100	100	20		73
SA-20		*passage*	*collapse*	*fringe*		
	conv	100	100	80		93
	mod	100	100	40		80
	clear	80	100	60		80
SA-36		*axes*	*accused*	*displays*		
	conv	20	40	40		33
	mod	60	80	80		73
	clear	40	80	100		73

4.2.3.2 SA-07

Like RG-37, SA-07 showed an intelligibility increase for clear/norm, but a decrease for modified conv/norm. Subjectively, the difference between conv/norm and the modified conv/norm utterance for this example was an amplification of the final keyword "debt" at the expense of slightly attenuating the rest of the sentence. Indeed, Table 4.1 shows a 40 point increase for this word. However, it appears that the attenuation of other words and segments resulted in an intelligibility degradation that outweighs this increase for the final keyword.

The entire drop in intelligibility for modified conv/norm can be attributed to the third keyword, "balances". This word has 100% intelligibility in the original conv/norm, but only 20% after segment normalization. Reviewing the listener responses, it was discovered that all mistakes on this word were due to substituting the word "bounces" for "balances". During the segments which distinguish these two words (i.e segments between /b/ and /n/), the energy redistribution results in a 20% decrease in RMS level, which could be responsible for the word substitutions. Because the RMS reduction of these segments also occurs in clear/norm, other changes must be present which allow this drop in power level without a drop in intelligibility. However, no major differences could be heard with subjective listening.

4.2.3.3 SA-20

Another sentence with unexpected results is SA-20. This sentence was originally chosen because of the large intelligibility increase of clear/norm over conv/norm seen in a previous study. However Figure 4-6 has shown that intelligibility was actually lower in clear/norm than in conv/norm with the current results. Table 4.1 reveals that the conv/norm score for SA-20 was 93%. Therefore, intelligibility scores for modified conv/norm and clear/norm are seen as a decrease despite both having received a high score of 80%.

This 93% score for conv/norm is substantially higher than the 38% average for the other six examples of conv/norm and therefore skews the statistics considerably.

If only the other six sentences are considered, average scores for conv/norm, modified conv/norm, and clear/norm become 38%, 52%, and 76%, respectively. For these numbers, energy redistribution accounts for about 35% of the difference between conv/norm and clear/norm, as opposed to the 27% reported when using all seven sentences.

The initial study on this data [2] placed intelligibility for SA-20 at 46% and 88% for conv/norm and clear/norm, respectively. This 47 point difference in score for the same conv/norm utterance highlights the inconsistencies between the two tests and suggests more complete, consistent testing must be done.

4.2.3.4 SA-36

Finally, sentence SA-36 is considered. Figure 4-6 showed that the segment normalized utterance had an intelligibility increase equal to that of the clear/norm utterance. This would suggest that segmental power levels are solely responsible for the advantages of clear speech in this case. It would be desirable to know why this occurred or if some segment or word can be singled out as primarily responsible for this behavior.

Table 4.1 shows that the intelligibility increase is evenly distributed between the three keywords. Inspection of listener responses did not reveal any consistent trend in keyword errors. Therefore, it is difficult to characterize what segments are responsible for this intelligibility increase. Subjectively, there appears to be very little difference in conv/norm before and after segment normalization despite the 40 point intelligibility increase measured in testing. With such subtle audible differences, the test results for this example are surprising. Averaging results from a larger number of listeners may be required for more accurate intelligibility scores.

4.3 Summary

Preliminary testing has shown that segmental power differences between clear/norm and conv/norm may account for about one-third of the overall intelligibility difference between these two modes of speech (for a select group of the "most successful"

examples). The sentence location dependencies of intelligibility differences between clear and conversational speech are almost entirely accounted for by segment level RMS values. However, due to the small test set and conflicting results with previous studies, further tests are needed to draw more meaningful conclusions.

Chapter 5

Discussion and Conclusions

When analyzing clear speech results, it is important to keep in mind the ultimate goal of creating a signal processing algorithm able to convert conversational speech into some reasonable approximation of clear speech (i.e. something with comparable intelligibility). This final chapter takes a step back to discuss some of the issues associated with such a transformation in the context of the information presented in this thesis.

5.1 Problem Formulation, Constraints, and Difficulties

To carefully approach any engineering task, it is important to carefully state the problem one is trying to solve and define the metrics one will use to gauge its success. Developing a successful transformation will depend highly on the constraints imposed, which in turn should be derived from the desired application.

Following the convention used by previous studies, the general objective explored here was to find a speech transformation algorithm that could increase intelligibility subject to the following constraints and definitions:

- Sentence-level RMS levels must be fixed
- Sentence-level durations must be kept approximately constant.

- Intelligibility of a sentence is defined as the percentage of keywords (nouns, adjectives, adverbs) correctly identified when presented in noise.

While these constraints have proven useful to investigate clear speech, variations may be preferred or required. The following is a discussion of other various properties and associated constraints to offer some perspective on what difficulties may arise with each.

5.1.1 Normal vs. Reduced Speaking Rates

To require a transformation to be at normal rates, the definition of "normal rates" must be clarified. In this and previous studies, "normal rates" was defined at the sentence level. However, we have seen that fairly significant deviations in duration occur at the levels of words and phones. Many of these deviations are large enough to create significant problems for systems requiring audio/visual synchronization (i.e. hearing aids). Therefore, for these applications, sentence level durations are likely not the best constraint for speaking rate.

5.1.2 Level Normalization

For a fair comparison of various utterances, it is clear that some normalization must occur. Sentence level power normalization is a natural choice, but it allows speakers to increase SNR in various parts of the utterance. This may be undesired if searching for acoustic features other than amplitude.

Ultimately, how normalization must be done will depend on the desired application. Many applications may require normalization to a maximum peak level instead of limiting sentence level power. With this constraint, the results in Chapter 4 could be quite different. For these applications, there already exist techniques to increase effective loudness under the constraint of maximum peak value [10, 4].

5.1.3 Knowledge of Speech Content

When humans create naturally produced clear speech, they do so with knowledge of the meaning of the words they are saying. It seems likely that this knowledge is used in the creation of clear speech. For example, the desired transformation may be required to treat keywords differently than non-keywords (as is suggested in Chapter 3). If so, knowledge of what is being said is needed and the solution becomes much more complex. Any system which does not utilize speech content will likely not be able to perform as well as a system that does (i.e. not as well as naturally produced clear speech).

5.1.4 Real-time vs. Non-Real time

Many useful applications of speech enhancement require the processing be done in approximately real-time, such as telecommunication systems and hearing aids. Computational limitations aside, this constraint limits our transformation to be a causal system (with some flexibility due to delays in human perception). If a causal or real-time constraint is imposed, some potentially useful modifications will not be possible. For example, we have seen that it may be possible to raise intelligibility by amplifying words near the end of a sentence. However, this requires knowledge of how close to the end of a sentence a given word is, and therefore it requires non-causal processing.

5.1.5 Retention of Original Speech Characteristics

Imagine a transformation that significantly raises the intelligibility of a given conversational utterance, but the output speech does not resemble speech of the original speaker, i.e. speaker identification information is lost or skewed. This could be completely unacceptable for some applications. Generally speaking, different applications may require the transformed signal to retain certain characteristics. For example, the conv/norm sentences for speaker RG in this study had a fairly high pitch with significant variation, leading to a "light and pleasant" quality to the utterance. After pitch modification, the pitch was lower and had less variation, which gave it a more

"serious" tone. While these differences were subtle in this case, some applications cannot allow such a change in prosodic effects, even if they result in slightly increased intelligibility.

5.1.6 Retention of Original Underlying Meaning

If emphasis of keywords (by increased power) truly does account for a substantial part of the high intelligibility increase, then in theory a system could attempt to do this automatically. While this may raise intelligibility, it may also cause some other unwanted problems in some cases. Some words that may need to be amplified in conversational speech to raise intelligibility may change the underlying message being conveyed. For example, a given sentence may have several interpretations depending on which words are stressed or emphasized. For some applications this fact will have little impact; however, these effects could play a role in some cases for some applications.

5.1.7 Speaker Independence

Any application would benefit from a speaker independent transformation if possible, and most would require it. However, as Krause discusses in detail, clear/norm and conv/norm acoustical differences appear to have a fairly significant dependence on speaker [3].

5.1.8 Intelligibility Metric

While the intelligibility metric used throughout this study may, in fact, be the best one to use, it is important to be aware that it has some limitations. With long-term normalization as a level constraint, a transformation which placed nearly all energy from non-keywords into keywords would constitute a valid solution (using the strict problem formulation discussed). This would substantially boost SNR during keywords and thus increase intelligibility while leaving non-keywords inaudible. To render this

as an unacceptable solution, either the intelligibility metric or level constraint of the problem formulation would need revision.

5.1.9 Summary

While this list is not exhaustive, it points out the fact that the modification of speech to increase intelligibility may have many different solutions and approaches, each with its own performance and all with substantial difficulties. Depending on the application, various constraints may need to be placed on an acceptable transformation, and adding constraints will likely decrease the maximum obtainable intelligibility improvement.

5.2 Ideas for Future Research

5.2.1 More Complete Testing of Segment Normalization

Results from Chapter 4 suggest that intelligibility of conv/norm can be increased by simply redistributing the energy of segments. Due to inconsistencies with previous studies, conducting more extensive intelligibility tests with more than five listeners and more than seven example sentences will be needed for further investigation. More precise information on what percentage of the advantages of clear speech can be attributed to power levels will be very helpful.

5.2.2 Attempts at Generalizing Energy Redistribution

Modification by energy redistribution has thus far been discussed assuming that the clear/norm utterance exists and its specific segmental energy properties are explicitly imposed onto the corresponding conv/norm utterance. It has been shown that such modifications can moderately improve intelligibility under the constraints proposed. Therefore, the next step would be attempting similar transformations without knowledge of the clear/norm utterance. To do so, generalizations of how energy is

distributed in clear/norm would have to be formalized. Such rules would likely include increasing energy in the later part of the sentence and increasing energy in keywords. Also, some rules could be tested to perform intra-word energy redistribution motivated by Figure 3-16, such as increasing energy in word-initial and word-final phones.

In addition to attempting to mimic energy distributions of clear speech, one could try to extend these trends even further than is done naturally. Perhaps more drastic redistributions of energy could result in intelligibility beyond that of naturally produced clear speech.

5.2.3 Other Modifications

While time-scale and fundamental frequency modifications were explored, they were not tested thoroughly in this study. If the quality of these transformations could be improved slightly, it is possible that they could offer a larger benefit in intelligibility than segment normalization alone. After subjective testing, it appears that non-uniform time-scale modification along with segmental power normalization will increase intelligibility if done properly. The effect of pitch modifications is more uncertain.

In general, various other modifications can be performed that aim to transform conv/norm to something sounding as similar to clear/norm as possible. With recordings of each sentence in both modes, these modifications can be done explicitly tailored to change some feature to match its characteristics in clear speech. Ideally, the intelligibility of the modified utterance would approach that of clear/norm as more relevant modifications are added. If this occurs, then the necessary properties of clear speech will be limited to those which were included as modifications. This approach of using custom transformations for each utterance provides the ability to test the effects of various features without requiring general rules of their behavior.

5.2.4 Collection of Additional Data

Additional data of clear speech at normal rates would be very useful. Studying additional speakers not only offers more data, but could help in distinguishing different speaker strategies and what separates them. Figure 3-1 showed that there is a large variation in intelligibility increase across sentences for a given speaker. If additional speakers are trained similarly and record the same set of sentences, it would be interesting to discover if certain sentences have similar intelligibility improvements across speakers. If so, what makes some sentences more susceptible to improvements than others could be explored.

5.2.5 Higher-Level Modifications

One of the reasons that a conversational-to-clear transformation is pursued is because there is an existence proof that it is possible: humans can naturally transform conversational speech into clear speech when presented with difficult hearing environments. However, this wording is somewhat misleading. Humans do not produce clear speech by modifying conversational speech; rather, the decision for clear speech is made before the speech production process begins. For a machine to perform such a transformation in an analogous way, it would have to be done at a higher level than conventional signal processing. Therefore, to achieve the creation of human-like clear speech from conversational speech, more drastic measures of processing would be needed.

This line of thinking could motivate an approach such as in Figure 5-1, where automatic speech recognition (ASR) is performed to add the ability for the transformation to be context-dependent. A system which knows what words or phones correspond to a waveform segment could perform higher level modifications which indeed may be very critical to intelligibility. Some transformations that could be possible (in theory) with such a system are insertion of stop bursts when omitted, insertion of schwa vowels after voiced plosives, and increasing of amplitude and/or duration during words or segments deemed crucial for intelligibility. How to classify

segments as "crucial for intelligibility" is an area for more research, but it could be based on keywords (i.e. nouns, adjectives, and adverbs), statistically less common words, or perhaps words containing complex articulatory movements.

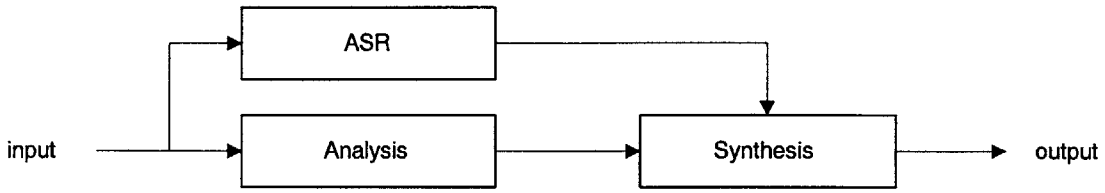

Figure 5-1: A higher level approach to producing clear speech from conversational speech which can utilize context depended modifications. Some form of Automatic Speech Recognition (ASR) is used to determine what processing is needed.

Due to the complexity and inaccuracies of even the best speech recognizers, a useful system based on this concept may not be realizable today. However, the ASR component could be replaced by human intervention for the purpose of researching clear speech. Also, the ASR system may not have to be as thorough or complicated as required by other applications. It may be sufficient to know some information about the signal (e.g. identifying keywords vs. non keywords, or identifying word and sentence boundaries) rather than requiring a full and correct transcription of the utterance.

5.3 Conclusions

This thesis has revealed several properties of clear speech at normal rates which should be considered when researching effective ways to use clear speech for intelligibility enhancement. Inspection of previous test results indicate that large portions of intelligibility improvements of clear speech are due to relatively small amounts of the data. The resulting "watering down" of the data may explain why previous acoustical analyses did not yield more striking differences between the two modes.

In looking specifically at "successful" examples of clear speech, it was observed that intelligibility differences are largely due to clear speech maintaining a fairly constant level of clarity, while conversational speech tends to decrease throughout

the course of a sentence. This decrease in clarity in conversational speech is due at least in-part to keywords toward the end of the sentence having less energy than in clear speech.

Preliminary tests have suggested that redistributing the segmental energy may be a necessary but insufficient step for transforming conversational speech to have the intelligibility characteristics of clear speech. Further studies will be needed to verify these results and test other modifications which explicitly force conversational sentences to resemble their clear speech counterparts. Ideally, as more modifications are added, the modified utterances will approach the clear speech utterances in terms of both aural similarity and measured intelligibility. If this is accomplished, an enhancement algorithm may be developed by generalizing these modifications so that they can be applied to conversational utterances for which clear speech is not available.

Bibliography

[1] F.R. Chen. Acoustic characteristics and intelligibility of clear and conversational speech. Master's thesis, Massachusetts Institute of Technology, 1980.

[2] J.C. Krause. The effects of speaking rate and speaking mode on intelligibility. Master's thesis, Massachusetts Institute of Technology, 1995.

[3] J.C. Krause. *Properties of Naturally Produced Clear Speech at Normal Rates and Implications for Intelligibility Enhancement.* PhD thesis, Massachusetts Institute of Technology, 2001.

[4] J. S. Lim. *Speech Enhancement.* Prentice Hall, 1983.

[5] 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*, 95:1581–1592, March 1994.

[6] 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.

[7] 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.

[8] 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.

[9] T. F. Quatieri and R. J. McAulay. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:744–754, August 1986.

[10] T. F. Quatieri and R. J. McAulay. Peak-to-rms reduction of speech based on sinusoidal analysis/synthesis. *IEEE Transactions on Signal Processing*, 39:273–288, February 1991.

[11] T. F. Quatieri and R. J. McAulay. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40:497–510, March 1992.

[12] 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*, 39:494–509, June 1996.