TOWARDS AN OBJECTIVE MEASURE OF SPEAKERS' INTELLIGIBILITY DERIVED FROM THE SPEECH WAVE ENVELOPE

by

DOROTHY C. HOEK

B.A., UNIVERSITY OF BRITISH COLUMBIA, 1985

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE SCHOOL OF AUDIOLOGY AND SPEECH SCIENCES

THE FACULTY OF MEDICINE

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

August, 1988

5 b

Department of Audiology & Speech Sciences

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date August 9, 1988

ABSTRACT

This study investigates the possibility of a relationship between amplitude modulation in the speech envelope and a speaker's intelligibility or articulatory clarity. It aims at developing an intelligibility measure called the Modulation Index (MI).

Speech samples from several English speakers and one French speaker were recorded and digitized. Speakers were asked to produce speech under three articulatory conditions: Underarticulated, Normally Articulated, and Overarticulated. A computer program was developed for calculation of MI, based on the amount of amplitude modulation depth in the envelope of each digitized speech sample. The MI values so obtained were compared with the corresponding ratings from English-speaking listeners who judged the articulatory clarity of the recorded utterances.

Results indicate that the relationship between the perceptual data and the Modulation Index in its present form is weak and non-monotonic. Several factors may have affected the results of the comparison between the MI values and the perceptual data. There are indications that speakers were not always successful in producing the intended articulatory conditions. Also, despite precautions, there were some differences in intensity and duration between utterances from the three conditions.

It is concluded that there is some correlation between amplitude modulation in speech envelopes and speakers' intelligibility or articulatory clarity. However, the Modulation Index will require modification before it can become a useful tool. Some modifications were briefly explored, and possible further modifications to both the Modulation Index and the experimental design are suggested for future investigations.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

LIST OF APPENDICES

ACKNOWLEDGEMENT

I would like to thank everyone who has had a part in this thesis. In particular, I would like to thank:

Dr. André-Pierre Benguerel for sharing a small part of his knowledge with me, and for his assistance throughout the preparation of this thesis.

Noelle Lamb for her encouragement and interest, and for serving on my committee.

My subjects for their kind cooperation.

My family and Mike for their patience and support.

My classmates, especially Charlotte, for their friendship.

CHAPTER ONE

INTRODUCTION

Intelligibility of speech refers to the ease with which speech may be understood by normal hearing listeners. Speech which is unintelligible may have been affected by interfering conditions, such as reverberation or noise, or it may have been unintelligible as it left the speaker's lips.

Various methods have been developed in attempts to measure speech intelligibility. Some of these, such as the NTID scale (cited in Doyle, 1987), are subjective and are based on listener ratings. Others, such as the Articulation Index (French & Steinberg, 1947), are objective measures based on acoustic analyses. Furthermore, some methods were designed or adapted for measurement of speech intelligibility at the source (Kondraske, 1985, for example), while others, including the Articulation Index, assume full intelligibility at the source and measure the degradation of signals over speech transmission systems.

Speech intelligibility measures will be discussed in detail in Chapter Two, but some general observations which motivated the present study are in order. Methods involving listener judgments have various disadvantages. The experience and biases of listeners affect their judgments, necessitating large sample groups, and increased analysis time (c.f. Doyle, 1987). In addition, while listeners may make gross judgments about the adequacy of the speech, they cannot make the fine acoustic analyses necessary for pinpointing the particular attributes of a speech signal which contribute to its relative intelligibility. Without knowledge of the aspects of the acoustic signal which are behind perceptual errors, attempts to remedy the problems in the system will at best be based on educated guesses. For example, if the source of poor speech intelligibility in an auditorium is known to be reverberation, engineers can improve the situation by placing acoustic tiles, or by other means specific to the problem. If, on the other hand, the factors detrimental to speech transmission in that auditorium are unknown, the solution to the problem can only be found by trial and error. Even those tests which identify perceptual errors affecting individual speech sounds do not reveal the acoustic correlates of those errors. Acoustic indices, in contrast, give consistent ratings from measurement to measurement, are less time consuming to score, and give some information as to the physical qualities contributing to the intelligibility of speech.

The purpose of the present study was to explore the possible application of the concept behind an acoustic index - the Modulation Transfer Function (MTF) - to an index of intelligibility at the source, i.e. the speaker. Previously, the use of the MTF concept has been confined to the study of speech transmission systems, or to psychoacoustic perceptual studies. The MTF reflects the reduction in modulation depth of an acoustic wave propagating from source to receiver when the intervening transmission system introduces reverberation, noise, and frequency filtering as contaminants (see Figure 1 for an illustration of the principle of modulation reduction). Houtgast and Steeneken (1973) compared MTF values obtained from various transmission systems to speech intelligibility measures obtained using Phonetically Balanced word scores. Finding a strong correlation between the two measures, the authors proceeded to incorporate the MTF as an integral part of the Speech Transmission Index (Steeneken & Houtgast 1980; Houtgast and Steeneken 1971, 1985). Houtgast and Steeneken (1973) found a near linear relationship between speech intelligibility scores and retention of amplitude modulation in the received signal as measured by the MTF. Thus, depth of modulation of a speech signal seems to correlate with speech intelligibility.

In the present study, the goal was the development of a rating tool for assessment of intelligibility as affected by a speaker's articulatory clarity. The scores obtained from a measure of modulation depth (the Modulation Index or MI) were compared to listener's perceptual judgments of articulatory clarity for recorded speech samples from several speakers.

Efforts were made to minimize contaminating variables which might affect the results. Background noise and

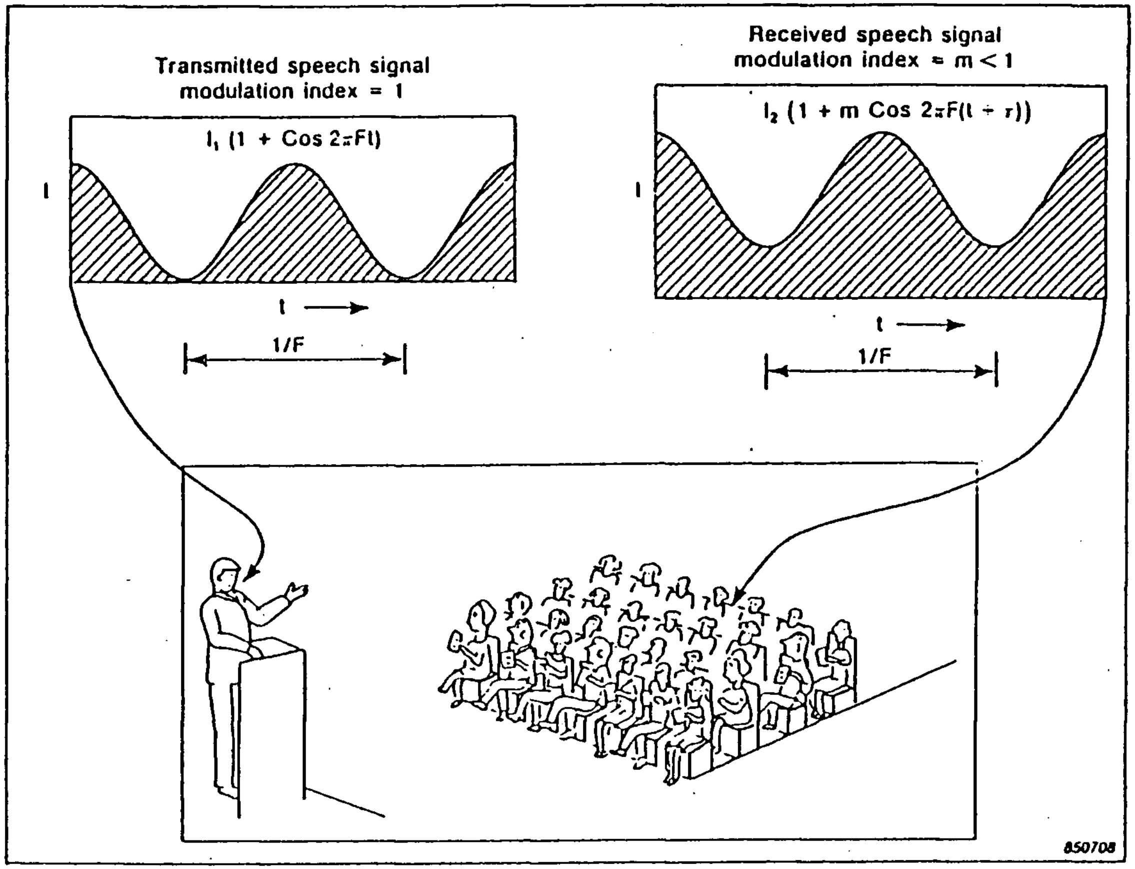

Transmitted speech signal
modulation index = 1
$I_1 (1 + \text{Cos } 2\pi Ft)$
I
t →
1/F
Received speech signal
modulation index = m < 1
$I_2 (1 + m \text{ Cos } 2\pi F(t - \tau))$
I
t →
1/F
850708

Figure 1. Illustration of the reduction in modulation of a speech signal caused by background noise and reverberation. (reproduced from Steeneken and Houtgast, 1985).

reverberation, which could both decrease the inherent amplitude modulation in the produced speech samples, were minimized by making all of the recordings in a sound proof booth with acoustic tiling on the walls and ceiling. High quality recording equipment was used to avoid frequency filtering, and other effects which would reduce the recording quality. Furthermore, in its present formulation, the new measure (MI) is sensitive to absolute intensities, and the effect on the measure of timing differences, whether in duration of samples or in spacing of words or duration of individual phonemes, is difficult to predict. Therefore, in order to minimize these effects, speakers practiced keeping intensity and rate of speech as constant as possible; they received feedback as to their performance from the experimenter, and they were urged to maintain this constancy during recording sessions.

These precautions were taken in order to maximize the influence of the variables of interest - amplitude modulation depth and articulatory clarity. Nevertheless, other contaminants, unrecognized and unaccounted for, may have been present. This is a hazard of undertaking an exploratory study using natural speech samples.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter is a survey of the available measures of speech intelligibility, and of related literature on the subject. Tests of speech recognition aimed at quantifying perception of speech by impaired listeners are not reviewed here. (These tests include tests of word discrimination such as the CID W-22 word lists, among others).

The study of speech intelligibility has many useful applications. Knowledge of the factors which degrade intelligibility and which are most influential in a particular setting can contribute to the design and construction of acoustically optimal auditoria, lecture halls, classrooms, telephones, public address systems, and hearing aids. Speech-language pathologists may employ these methods when making diagnoses or assessing progress. Thus, the potential applications of speech intelligibility measures are many and, not surprisingly, there have been many different measures developed.

2.2 FACTORS AFFECTING SPEECH INTELLIGIBILITY

2.21 FACTORS INVOLVING THE SPEAKER

Several studies of the acoustic variables which affect intelligibility of an individual's speech are to be found in the literature. An exhaustive list of these studies is not necessary for this survey, but a description of particularly relevant studies will illustrate to the reader some of the approaches which have been explored. Monsen (1978) studied the speech of hearing impaired children. Three acoustic variables accounted for 73% of the variance in normal listeners' judgments of the children's intelligibility. These differences were: (1) the differences in voice onset time between /t/ and /d/ (accounted for 48.5% of the variance); (2) the second formant difference between /i/ and /ɔ/ (accounted for 20.5% of the variance); and (3) the presence (normal) or absence of rapid spectral change between a syllable initial liquid or nasal and the following vowel. The other phonetic variables found to contribute little or not at all to intelligibility were voice onset time differences between /k/ and /g/ and between /p/ and /b/, first formant differences in vowels, and extent of the second formant frequency change in the diphthong /ai/. In addition, Monsen cited previous studies in which the authors claimed that variables affecting acoustic prosodic parameters such as duration, rate, and fundamental frequency (Voelker, 1938; Hudgins and Numbers, 1942; Hudgins, 1960; John and Howarth, 1965; Brannon, 1966; Ando and Canter, 1969) contribute significantly to the relative intelligibility of speech. These findings, however, were not confirmed by Monsen (1978).

Approaching the subject differently, Ananthapadmanabha (1983) regards the speech signal as the convolution of the source (glottis), and the vocal tract filter, after Fant's (1960) model of speech production. The parameters of speech originating at the source, collectively designated as "source dynamics", considered by Ananthapadmanabha, are voicing and plosive contrasts, intensity changes, and natural pitch variations. Formant information is imposed on the source dynamics by the vocal tract filter.

In order to isolate the source dynamics and exclude formant information, Ananthapadmanabha passed a speech signal through a so-called "epoch" filter. The input to this kind of filter is voiced speech while the output consists of pulses corresponding to each peak of vocal source excitation in the signal. Each peak is termed an epoch. The signal is first passed through a third-octave band-pass filter centred at 4 kHz, then rectified and low pass filtered with a 340 Hz cutoff frequency.

Even without the formant information, enough phonetic information for the comprehension of speech remained. As a result, Ananthapadmanabha concluded that source dynamics has a much stronger role to play in the perception of the phonetic information than previously believed.

When taken together with Monsen's (1978) results, Ananthapadmanabha's conclusion is difficult to evaluate, as the variable groupings by the two authors overlap. Since source dynamics included Monsen's most important variable (consonant voicing contrasts, or voice onset time differences), Monsen would have predicted that intelligibility would be maintained in Ananthapadmanabha's processed speech. On the other hand, Monsen would have incorrectly predicted loss of intelligibility when Ananthapadmanabha excluded formant information, which Monsen considered important.

Metz, Sama, Schiavetti, Sitler, and Whitehead (1985) also investigated factors affecting intelligibility of hearing impaired speakers. These authors replicated Monsen's study and confirmed his major findings. They agree with Monsen in labeling segmental information as the primary dimension of speech intelligibility. Contrary to Monsen, however, and in concert with previous authors, Metz et al. suggest that prosodic features are an important secondary dimension.

The studies mentioned to this point have dealt with inter-speaker differences in intelligibility. Picheny, Durlach and Braida (1985, 1986) investigated instead differences in intelligibility of speech produced by the same speakers in different situations. In particular, these studies focused on the acoustic characteristics of "clear" speech. Clear speech was defined as speech intended for hearing impaired listeners or produced in noisy environments. The authors contrasted clear speech with "conversational" speech, the latter being speech intended for normal hearing listeners in the absence of competing noise. In their 1985 study, Picheny et al. found the following differences between clear and "conversational" speech: (1) the duration of sentences produced in clear speech almost twice the duration of sentences produced in conversational speech, and this difference was a reflection of both additional pauses and increased duration of individual speech sounds in the clear condition; (2) there were more instances of vowel reduction and consonant deletion in conversational speech than in clear speech; and (3) there were differences between the conditions in the short-term spectra of individual speech sounds; for example, in clear speech, consonant intensities tended to be greater in relation to neighboring vowels than in conversational speech. As a cautionary note, however, the results of this study may have been affected by some of the preliminary instructions given to speakers before recording of the "clear" condition, as the following excerpt from Picheny et al. (1985) indicates:

> "The talkers were also told to enunciate consonants more carefully and with greater (vocal) effort than in conversational speech and to avoid slurring the words together." (p. 97)

These instruction probably introduced a disproportionate number of pauses and other artifacts into the speech samples.

Picheny et al. (1986) confirmed their 1985 results, and added findings that the long-term RMS (root mean square) spectrum level was not substantially different between clear and conversational speech, and that there was a wider range of fundamental frequencies in clear than in conversational speech. Again, Picheny et al.'s work supports Metz et al.'s (1985) view that prosodic features play an important part in the relative intelligibility of speech.

2.22 FACTORS INVOLVING THE TRANSMISSION SYSTEM

French and Steinberg (1947) listed a number of factors which can affect intelligibility. These include the intensity of the signal, background noise in the system, reverberation, and phase distortion. Miller and Nicely (1955) added low- and high-pass filtering to the list, but pointed out that low-pass filtering, in its effect on speech, is roughly equivalent to the effect of background noise because of the lower intensity of the high frequency components of speech. Given the nature of the majority of sensorineural hearing losses (i.e. with the greatest degree of damage to high frequency hearing), the low-pass filtering effect of low fidelity audio systems, and the prevalence of background noise as a barrier to speech transmission, low-pass, rather than high-pass filtering is more likely to be an important factor in determining speech intelligibility in situations outside acoustic laboratories.

Miller and Nicely defined five "articulatory" dimensions in speech. These were voicing, nasality, affrication, duration, and place of articulation. Voicing and nasality were found to be most robust when speech was subjected to noise or low-pass filtering, whereas place of articulation was the most easily disrupted dimension under these conditions. None of the dimensions was particularly resistant to high-pass filtering, since in this case most of the acoustic energy in the consonants was removed, leaving the remaining available information at very low intensity, and consequently inaudible for the purposes of speech perception.

Intelligibility of speech is also adversely affected by reverberation. Reverberation is defined by Rettinger (1968) as:

> "sound persistence due to repeated boundary reflections after the source of sound has stopped." (p. 85)

Boundaries in this case are surfaces such as walls or ceilings, or any object in an enclosed space. Speech intelligibility is reduced by reverberation because persisting sound energy results in overlap of successive speech sounds and blurring of the signal.

The determining factor in the susceptibility of speech to degradation through reverberation has been traditionally identified as the reverberation time. Morse and Ingard (1968) define reverberation time as:

> "The length of time it takes the mean energy of the wave to reduce to a millionth part of its initial mean value". (p. 558),

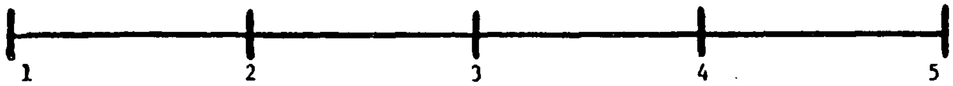
or, in other words, the time taken for the wave energy to decrease by 60 decibels. Lochner and Burger (1964) concluded that speech is unaffected by reverberation only at reverberation times below 0.3 seconds, but that the signal and its reflections are partially integrated at times between 0.3 and 0.8 seconds. Morse and Ingard add that if the speech signal changes significantly in a time less than one tenth of the reverberation time, the original signal will be blurred by reflected sound energy. Furthermore, Crum (1974) stated that a reverberation time of 1.2 seconds or more decreased intelligibility for normal hearing adults in quiet, and that the combination of background noise and reverberation reduced speech recognition performance more than predicted from the sum of the effects of the individual variables.

2.3 SPEECH INTELLIGIBILITY MEASURES BASED ON LISTENER JUDGMENTS

2.31 INTELLIGIBILITY SCALES

Perhaps the most subjective measures of intelligibility are intelligibility scales such as the two employed by Doyle (1987), which may be used, for example, for screening a population of speakers for intelligibility deficits. Figure 2 and Figure 3 describe these scales, which give no details as to the factors affecting relative intelligibility, but are quick to score and administer.

Doyle studied the use of these scales by audiologists assessing hearing impaired children's speech. The results indicated good intra-rater reliability, but poor inter-rater reliability, particularly for the scores assigned to the speech of certain individuals. Thus, listener bias is a concern in the use of intelligibility scales.



1	2	3	4	5
Speech is completely unintelligible.	Speech is very difficult to understand, only isolated words or phrases are intelligible.	With difficulty the listener can understand about half of the message (intelligibility may improve after a listening period).	Speech is intelligible with the exception of a few words and phrases.	Speech is completely intelligible.

Figure 2. The National Technical Institute for the Deaf (NTID) Scale of Intelligibility (after Doyle, 1987).

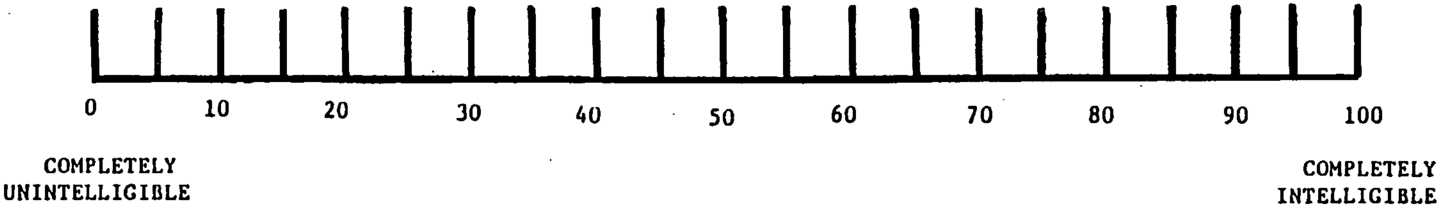


Figure 3. A percent rating scale of intelligibility (Doyle, 1987).

2.32 DETAILED JUDGMENT-BASED INTELLIGIBILITY TESTS

Black (1957) reviewed early intelligibility tests (see Appendix A for a list of the tests) and he advocated the use of a multiple choice format with answer forms provided. For example, a listener would be presented with a choice of four possible words, and he would circle the word he thought he heard. He compared the multiple choice format to tests where listeners simply wrote down the words they heard. The multiple choice tests had the advantage of reducing the burden on the phonetic knowledge of the scorer and the scoring time, but otherwise they were no more reliable (nor necessarily more valid) than the write-down tests.

Williams and Hecker (1968) compared the results of four tests aimed at the assessment of speech transmission systems (listed in Appendix A). These authors used various types of speech distortion (additive speech-shaped noise, peak clipping, and vocoding), and two different speakers for the test conditions. They confirmed the earlier finding of Hirsh et al. (1954) that individual intelligibility test scores varied relative to one another depending on the type of speech distortion introduced into the same transmission system. Furthermore, the scores obtained for their two speakers were more similar for some distortion types than for others.

More recently, Newman (1979) wrote the introduction to a chapter consisting of a collection of reviews of articulation tests used by speech-language pathologists. The tests are listed in Appendix A. These tests were developed to replace spontaneous speech samples with the idea of decreasing testing and analysis times, and ensuring that all phonemes were sampled in a given session. In contrast to intelligibility scales and the intelligibility tests discussed by Black (1957), articulation tests, as well as phonological analyses performed on spontaneous speech samples, give attention to the specific phoneme confusions which adversely affect intelligibility. Still, these tests provide only superficial descriptive information, since even the ear of a well trained phonetician cannot analyze phoneme errors acoustically. To summarize, Newman criticized these tests for their questionable validity and reliability, although he acknowledged recent efforts to improve this situation. In addition, he applauded the addition of suprasegmental phonemes to the content of some tests, which should help to improve test validity.

2.4 ACOUSTIC INTELLIGIBILITY MEASURES

2.41 THE ARTICULATION INDEX (AI)

The Articulation Index (AI) was conceived by French and Steinberg (1947). Although originally intended for assessment of telephone systems, the AI has been revised for several purposes, and it is well enough established to be described in an American National Standards Institute standard (ANSI 1969).

The Articulation Index was an innovation in that it summarized speech intelligibility into one number, independent of listener judgments, and was firmly based on acoustics. As originally formulated by French and Steinberg (1947), the Articulation Index is a ratio obtained by comparing an ideal input speech spectrum (empirically determined) with the actual output of a speech transmission system considering the effects of noise and band-pass filtering. The speech spectrum is divided into twenty frequency bands in the range 200-6100 Hz, each of which is considered to contribute equally to speech intelligibility. The use of twenty frequency bands provides a measure with good frequency resolution, useful in situations where sharp filtering of the speech signal takes place, or where there is narrow band or frequency specific background interference. However, later authors (eg. Kryter, 1962a; ANSI 1969; Humes, Dirks, Bell, Ahlstrom, and Kincaid, 1986) have used 1/3 octave bands or octave bands, together with weighting factors, with similar AI scores obtained.

Pavlovic (1987) provided a summary of the various modifications which have been applied to the AI since 1947, as well as updated tables of variables for AI calculations. The Articulation Index is calculated by means of the following two equations:

$$A = P \sum I_i W_i$$

$$s = T(A),$$

where A is the Articulation Index, P is the proficiency function (a measure taking into account the enunciation of the speaker and the familiarity of the listener with the materials), i is the frequency band under consideration, W_i is the proportion of the speech dynamic range within frequency band i which contributes to overall speech intelligibility over the transmission system, I_i is the ideal contribution of that frequency band, and s is speech intelligibility related to the AI through the empirical transfer function given in the second equation.

The original Articulation Index contained no provision for reverberant room conditions. Kryter (1962a) partly remedied this situation by providing correction factors based on reverberation time which, in modified form, were incorporated into the ANSI standard. However, Humes et al. (1986) found that these corrections were inadequate at signal-to-noise ratios worse than zero decibels, and Kryter himself stressed that the corrections were based on the results of a single study.

The validity of the AI has been the subject of numerous investigations. French and Steinberg (1947) provided a chart of AI scores compared to listener judgment scores obtained using a variety of speech materials (see Figure 4). As Kryter (1962a,b) pointed out, greater semantic redundancy of the speech materials results in a smaller AI score for any given intelligibility score. By semantic redundancy, he meant material in which meaning can be gleaned from syntactic cues, or in which a few words are repeated often, which allows subjects to guess at words more accurately. This

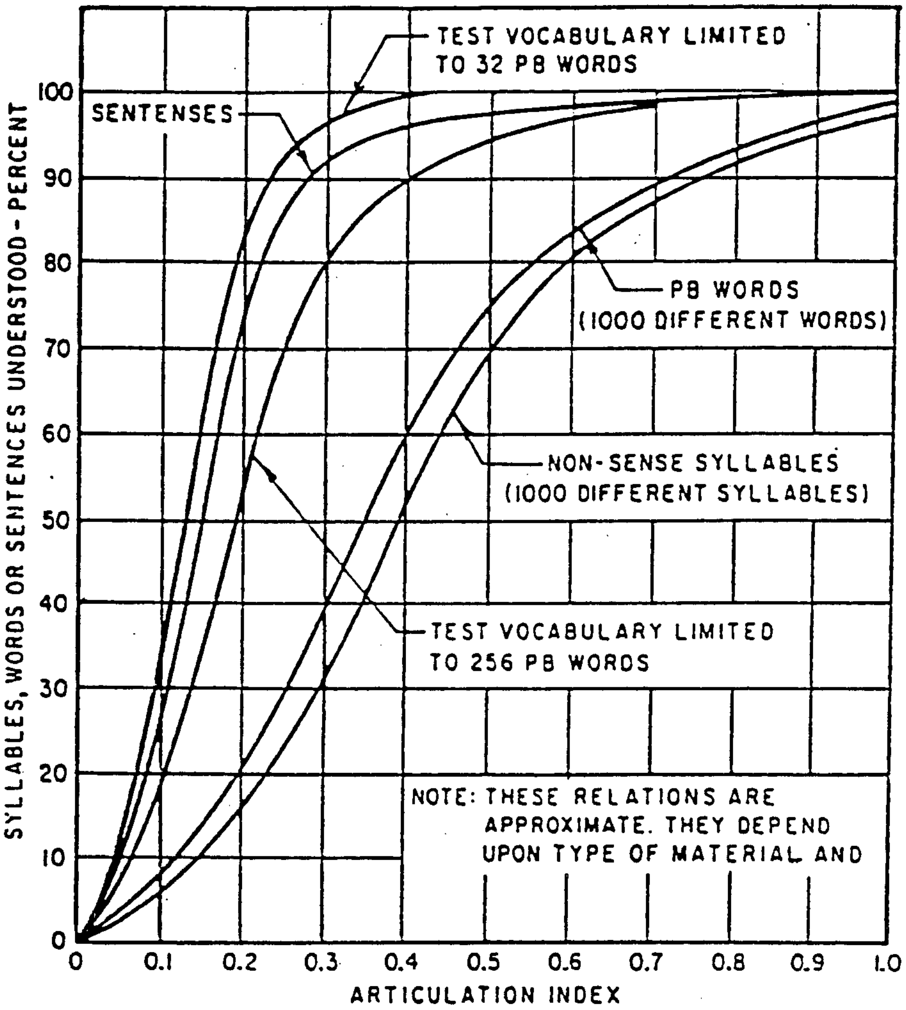


Figure 4. Relation between Articulation Index scores and PB-word scores. (reproduced from French and Steinberg, 1947).

means that an AI score has limited meaning in isolation from any specification of the speech material to which it applies.

Pavlovic (1984), Kamm, Dirks and Bell (1985), and Dirks, Bell, Rossman and Kincaid (1986) have investigated the validity of the AI when applied to hearing impaired listeners. They found that the AI was a good predictor of the performance of most listeners, with some exceptions among subjects with severe high frequency sloping sensorineural hearing losses.

2.42 THE SPEECH TRANSMISSION INDEX (STI, RASTI)

Houtgast and Steeneken (1971) introduced an acoustic index having one big advantage over the Articulation Index: provisions for taking into account peak clipping, band-pass filtering, and reverberation, as well as background noise, were built in, rather than added on as clumsy correction factors. Instead of natural speech signals, the STI employed artificial signals - another difference from the Articulation Index.

The original Speech Transmission Index (STI) calculates a weighted sum of spectrum differences between the two levels of an alternating signal in each of the five octave bands centred at frequencies ranging from 250 to 4000 Hz. The signal level difference at the input to a speech transmission system was compared to the difference at the output. This principle, as applied to background noise, is illustrated in Figure 5. The two levels of the alternating signal are referred to as two separate signals in the discussion which follows.

At the input, the two signals consisted of noise shaped to resemble the average long-term speech spectrum. One of the signals (Sound 1) was more intense than the other (Sound 2), but since the signals had the same spectral shape at the input, the intensity difference between them at the input was equal across the frequency range. If, however, the signals were passed through a transmission system containing background noise approximately equal in intensity to Sound 2, their intensity level difference ($\Delta L'$) at the output would be changed. Sound 2, combined with the noise, would result in an output signal (Sound 3) having higher intensity and a spectral shape different from that of Sound 2. Sound 1, however, being significantly more intense than the background noise, would be essentially unaffected by it. Comparison of the spectrum levels of Sound 3 and Sound 1 at the output, in each of the octave bands, would reveal the effects of the background noise, since the intensity differences between the two signals would no longer be equal across the frequency range. A weighted sum of these differences would take into account the different contribution of each octave band.

If reverberation was present in a system, the alternation rate (3 Hz at input) would be changed after transmission.

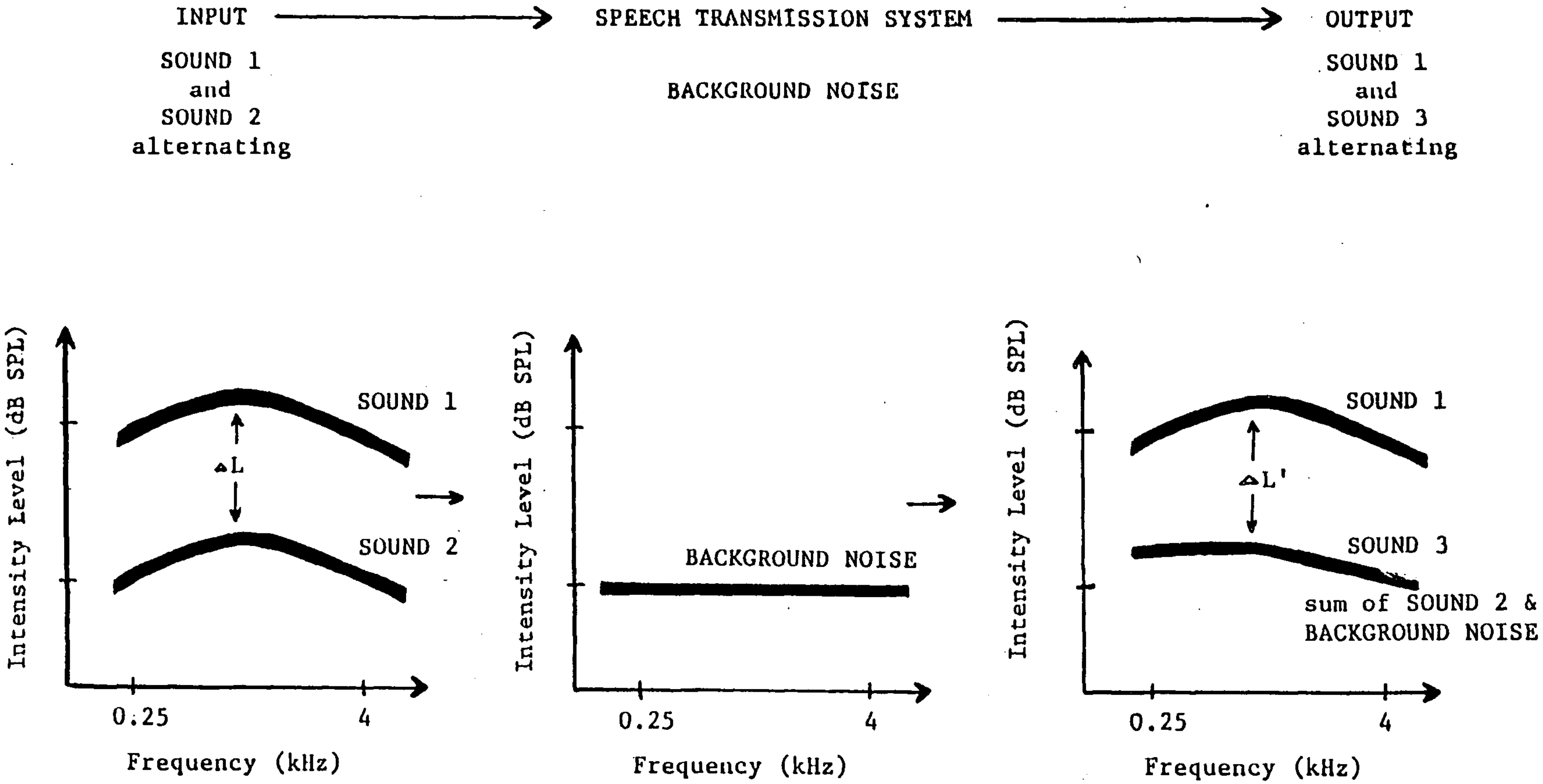

INPUT
SOUND 1 and SOUND 2 alternating
SPEECH TRANSMISSION SYSTEM
BACKGROUND NOISE
OUTPUT
SOUND 1 and SOUND 3 alternating
Intensity Level (dB SPL)
SOUND 1
ΔL
SOUND 2
0.25
4
Frequency (kHz)
BACKGROUND NOISE
ΔL'
SOUND 3
sum of SOUND 2 & BACKGROUND NOISE

Figure 5. The way in which the original Speech Transmission Index accounted for the effects of background noise.

Reverberation was measured by quantifying the change in alternation rate from input to output.

The signals, and the analysis procedure which followed, were developed and modified on the basis of comparisons to Phonetically Balanced (PB) word scores measured over fifty transmission channels. Various degrees and combinations of reverberation, band-pass filtering, interfering noise, and peak clipping were used as contaminants in the channels.

The formula for the STI in this early form was:

$$STI = 1/5 \sum_{i=1}^{5} \left(\frac{\Delta L'}{20 dB} \right)$$

where i is the octave band index, is the output intensity level difference (which incorporates alternation rate differences, if any), 20 dB is the initial intensity level difference between the two signals. Limitations indicated by the authors included inability to account for frequency distortion, center clipping, or extremes of intensity.

The next step in the evolution of the Speech Transmission Index was the incorporation of the Modulation Transfer Function (MTF) (Houtgast & Steeneken, 1973; Steeneken and Houtgast, 1980). The MTF concept, whereby the fluctuations in the envelope of an input signal are smoothed in the output signal by the effects of reverberation and background noise, originated in studies of visual perception (see Figure 1). In the revised STI, the influence of Modulation Transfer Functions for seven frequency bands of a sinusoidally modulated input signal of shaped pink noise are combined into a single score. The calculation scheme was adapted from the Articulation Index. The correlation of revised STI scores with (Dutch) Phonetically Balanced word scores is illustrated in Figure 6. Non-linear distortion, changing background noise during measurements, and frequency dependent reverberation times will result in invalid STI scores. A review of the development and applications of the Speech Transmission Index is presented in Houtgast and Steeneken (1985).

The authors have moved in three directions with the Speech Transmission Index.

The first is the development of measurement devices. Steeneken & Agterhuis (1978) described an STI meter used in field studies. More recently, (Houtgast & Steeneken, 1984; Steeneken & Houtgast, 1985, Dareham, 1986) RASTI (Rapid Speech Transmission Index) was introduced. This is a screening meter, in which only two of the octave frequency bands (centered at 500 Hz and 2000 Hz) are included, but the method of calculation is otherwise similar to the Speech Transmission Index.

The second direction is the design of acoustically optimal auditoria, using a desired STI value as a starting point (Houtgast, Steeneken and Plomp (1980) and Plomp, Houtgast and Steeneken (1980)). The specifications given were for the volume of the room, the reverberation time, the

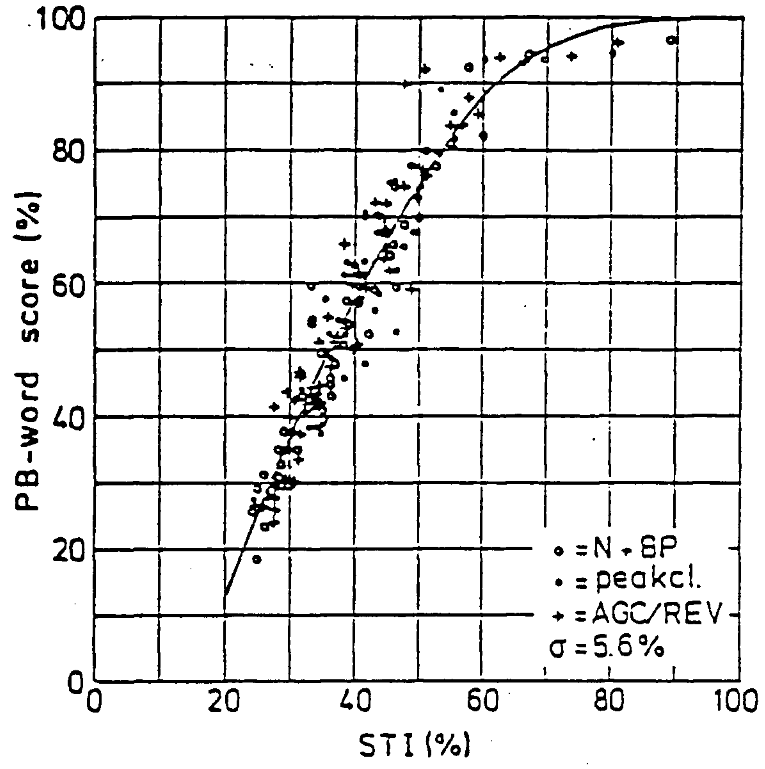


Figure 6. Relation between STI and PB-word score (Dutch words) for the conditions with noise, bandpass limiting, peak clipping, automatic gain control, and reverberation. The curve represents the best-fitting curve for all these data points. (Reproduced from Steeneken and Houtgast, 1980).

ambient noise level, the original signal intensity, and the distance between speaker and listener.

The third new direction for the STI is a computer ray-tracing model designed to provide an STI score for each individual audience position rather than merely one score for the entire auditorium (van Reitschote, Houtgast and Steeneken, 1981, 1983). In this model, the speaker is simulated by a point source which emits a signal to each audience position.

2.43 MODIFIED SPEECH TRANSMISSION INDEX (mSTI)

The mSTI is a hybrid of the AI and the STI in that it combines the modulation transfer function approach with Articulation Index weighting factors for one third octave band frequency analysis. Humes et al. (1986) created the mSTI as an improvement on its two predecessors for prediction of speech intelligibility performance by normal and hearing impaired listeners when speech was temporally and spectrally distorted. The mSTI did indeed prove superior to the AI and to the STI. The mSTI scores matched best with scores obtained by hearing impaired listeners on a speech recognition test.

2.44 ARTICULATION LOSS OF CONSONANTS (ALcons)

Peutz (1971) and Klein (1971) developed a measure called the Articulation Loss of Consonants measure (ALcons), so called because Peutz regarded the degradation of intelligibility over a transmission system as a loss of information, and because the measure proved to be much more consistent for transmission of consonants than of vowels.

Like the STI, ALcons directly accounted for reverberation effects, but, like the AI, it employed natural speech as an input signal. Peutz suggested that up to a certain critical distance (d_c), speech intelligibility is partially dependent on the speaker to listener distance. Above d_c, intelligibility is independent of distance, and varies with the reverberation time of the room. The following equations are used to obtain the ALcons measure:

$$\text{for } d < d_c, \quad \text{ALcons} = 200\,\frac{d^2T^2}{V} + a\ (\%),$$

$$\text{for } d \geq d_c, \quad \text{ALcons} = 9T + a\ (\%),$$

$$\text{with } d_c = (0.2\ s^{1/2}m^{2/3})\sqrt{V/T}$$

In these equations, d is the critical distance in meters, d is the distance to the listener in meters, V is the room volume in m , T is reverberation time (at 1400 Hz) in seconds, ALcons is the intelligibility score, and a is a correction for the skills of the listener, as measured by a speech recognition test. A modification is made to the measure if there is competing background noise in the room.

Peutz (1971) used very small groups of listeners (five to ten people) when validating the ALcons. Still, according to Lundin (1982), even though its validity is not well established, this measure is widely accepted.

2.45 DIRECT-TO-REVERBERANT INTENSITY METHOD (SRR)

The Direct-to-Reverberant Intensity method, better known as SRR, for Signal to Reverberation Ratio (Lundin, 1986), combines features of the ALcons with a frequency band approach similar to the Articulation Index. It was formulated to incorporate the built-in consideration of reverberation and simplicity of the ALcons with the frequency specificity of the AI. The SRR may be calculated according to the following formulae:

$$\mathrm{SRR} = -20.\log d/r_r$$

$$r_r = \sqrt{QA/16\pi} = 0.057\,\frac{(s)^{1/2}}{(m)^{1/2}}\sqrt{QV/\alpha\,T}$$

where d is the distance in meters between the source and the listener, r_r is the reverberation radius in meters (defined as the distance between the source of the original signal and the point where the original signal and the reverberant signal are equally intense (SRR = 0 dB)), Q is the directivity of the source (which corresponds to the proportion of sound from the source which actually reaches the listeners when the rest of the sound energy is dissipated around the room), A is the absorption of the room in metric sabins[1], V is the volume of the room in m^3, and T the is reverberation time of the room in seconds. The two formulae for calculation of r_r are related via

[1] The unit of absorption, the sabin, is named in honor of W.C. Sabine, and has the dimensions of one square foot. A metric sabin has the dimensions of one square meter, and is therefore equal to 10.76 sabins, since there are 10.76 square feet in one square meter.

Sabine's formula:

$$A = 0.163\ (s/m)\ V/\alpha T$$

where the absorption coefficient, α, is assumed to be equal to one.

Lundin (1986) found, however, that the SRR failed to provide better predictions of intelligibility than the ALcons, the AI, or the STI. He concluded that all four measures gave roughly equivalent predictions of the performance of normal hearing listeners in adverse conditions, but that these predictions overestimated intelligibility as measured by listener judgments.

2.46 SPEECH COMMUNICATION INDEX (SCI)

The SCI (Kryter and Ball, 1964) has not been widely adopted because its use requires sophisticated instrumentation. An intelligibility score is calculated on the basis of the signal-to-noise ratio in nine frequency bands, frequency shift, and peak clipping in the system under study.

2.47 PATTERN CORRESPONDANCE INDEX (PCI)

The PCI (Licklider, Bisberg and Schwartzlander, 1959) is another index which has only specific applications in systems analysis because of the complex instrumentation required. In this case, the pattern of the running power spectrum of a real speech signal is compared before and after passage through a transmission system. This method inspired Houtgast and

Steeneken (1971) when they originally created the Speech Transmission Index.

2.48 KONDRASKE'S METHOD

A recent attempt to quantify intelligibility of speech at its source is Kondraske's (1985) measure. He intends this method, which is still in its infancy, to be used by speech clinicians for assessment of patients with disorders such as dysarthria. A microphone is connected to a microcomputer which digitizes numerals spoken by the patient. The measure considers peak amplitude, average amplitude, peak to average amplitude ratio, inter-syllable time, and speed of articulation measured as the number of syllables produced in ten seconds.

2.49 MONSEN'S FORMULA

Having identified a small number of especially influential variables in the determination of speech intelligibility, Monsen (1978) (discussed above in section 2.21) developed the following formula to be applied to the speech of the hearing impaired:

$$I = 0.91(T_t - T_d) + 0.0214(F_i - F_{\text{ɔ}}) + 4.78(L,N) + 54.57,$$

where I is the index of intelligibility, T_t is the mean voice onset time of /t/, T_d is the mean voice onset time of /d/, F_i is the mean second formant frequency for /i/, $F_{\text{ɔ}}$ is the mean second formant frequency for /ɔ/, L and N are numerical variables reflecting the presence or absence of rapid spectral change following syllable initial liquids and nasals, and 54.57 is an empirically determined constant. Monsen tested the validity of his formula, and found a correlation of 0.86 between predicted and obtained intelligibility scores assigned by normal hearing listeners. This formula has not been widely adopted elsewhere, however, probably in part due to the time consuming spectrographic measurements required.

2.5 COMPARISONS OF ACOUSTIC INDICES

Some of the indices described (Monsen's, PCI, SCI) have limited applications and have not gained popularity since their introduction. The applicability of Kondraske's method, which has the promise of being available to clinical speech-language pathologists through office microcomputers, has yet to be determined. But what of the mSTI, the STI, the AI, the ALcons, and the SRR?

None of these measures is equipped to deal with non-linear frequency or amplitude distortion, or with extremes of signal intensity. The STI, mSTI, ALcons, and SRR are superior to the AI for reverberant conditions, but the ALcons and the SRR require an external correction in the presence of interfering noise.

Humes et al. (1986) found the mSTI to be superior to both the AI or the STI in prediction of intelligibility scores in the presence of temporal and spectral distortion. However, they also found that all three of these measures tended to underestimate loss of intelligibility in some hearing impaired subjects, a finding which is in agreement with those of Kamm et al. (1986) and Pavlovic (1984). Similarly, Lundin (1986) found that the AI, the STI, the SRR, and the ALcons all predicted higher intelligibility scores than those obtained through listener judgments.

2.6 FURTHER DEVELOPMENTS OF THE MODULATION TRANSFER FUNCTION

The Modulation Transfer Function has aroused the interest of others besides Steeneken, Houtgast and their colleagues. In 1981, for instance, Schroeder described the Complex Modulation Transfer Function (CMTF), which involves the use of Fourier transforms, and includes consideration of phase differences together with reduction in modulation depth in its calculation.

Elsewhere, Ahlstrom and his colleagues (Ahlstrom & Humes, 1983, 1985; Ahlstrom, Boney & Humes, 1985) have developed a method for assessing psychoacoustic MTFs by obtaining behavioural thresholds for temporal probe tones (tone pips at peaks or valleys of sinusoidally modulated speech noise). They have investigated Modulation Transfer Functions in subjects with normal hearing and sensorineural hearing losses, and subjects using compression and non-compression hearing aids.

2.7 CONCLUSION

Even given all of these limitations, the Modulation Transfer Function seems to be the most versatile of the measures on which indices have been based. It can account for both reverberation and background noise effects without external corrections, and it has warranted the attention of many authors working in several different directions, all with promising results. Perhaps in a form yet to be determined, the Modulation Transfer Function may well be the intelligibility measurement tool of the future.

CHAPTER THREE

METHODS AND MATERIALS

3.1 OVERVIEW OF THE EXPERIMENTAL DESIGN

The objective of this investigation was to explore the possibility of devising an acoustic measure of speech intelligibility when it depends only on the articulatory clarity of the speaker. The merit of this computed measure (henceforth referred to as Modulation Index, or MI), was evaluated by a comparison to listeners' perceptual judgments of the same speech materials.

For the purposes of this experiment, the range of articulatory clarity was divided into three "articulatory conditions". At the low end of the range, there was the "Underarticulated" (U) or mumbled condition, which was intended to correspond to poor intelligibility. In the middle of the range, there was the "Normally Articulated" (N) condition. At the top of the range, there was the "Overarticulated" (O) condition. This condition corresponded to maximally intelligible speech, such as that intended for hard of hearing listeners in noisy conditions.

Speakers producing sentence-length utterances were recorded. They were asked to produce the sentences in each of the three conditions mentioned. In most cases, but not all, the intended level of articulatory clarity was attained. The performance of the speakers will be discussed in detail in Section 4.12. MI values calculated for the speech samples were compared with the perceptual data, which was quantified in the form of listener judgments of articulatory clarity.

3.2 PREPARATION OF THE SPEECH SAMPLES

3.21 SPEECH MATERIALS

Initially, nine English and nine French sentences were composed. Each sentence contained nine syllables, and for each language there were three sentences containing predominantly labial consonants, three containing predominantly alveolar and palatal consonants, and three containing predominantly velar consonants. An effort was made to represent as many French and English phonemes in the sentences as possible.

3.22 SPEAKERS

Ten English speakers and one French speaker were recorded. Among the English speakers, five were male and five were female. The French speaker was male. Eight of the English speakers (four males and four females) were long-time or native Western Canadian residents who spoke the standard Western Canadian dialect. One male English speaker had a British (Received Pronunciation) accent, and one female speaker had a Newfoundland accent. The French speaker was a native of Lausanne, Switzerland; he also spoke English, but, unlike the other ten speakers, he recorded French sentences. All of the speakers were judged to have normal speech.

3.23 RECORDING OF SPEECH SAMPLES

Speech samples were recorded in a sound proof booth with acoustic tiling, using a Scully model 280 tape recorder and an AKG D202 dynamic microphone. Each speaker produced nine sentences under each articulatory condition, i.e. Underarticulated (U condition), Normally Articulated (N condition), and Overarticulated (O condition). The order of recording was Normal, Overarticulated, Normal, Underarticulated. The second Normal condition was used to enable the speaker to get back to his/her baseline after the Overarticulated condition, in preparation for recording the Underarticulated condition. Utterances in the second Normal condition were not used in MI calculations, or in the Listening test. The Underarticulated condition was recorded last because the experimenters felt that it would be the most difficult condition to produce intentionally, and that recording the other conditions first would possibly help the speaker form an idea of what was wanted.

The nature of the Overarticulated and Underarticulated conditions was not explained prior to recording the first condition (Normal). This was done because the Normal condition was intended to reflect the natural articulatory patterns of the speaker, and anticipation of the other conditions might have resulted in articulatory changes. Instructions for the Normal condition were to simply read the sentence through, without further prompting. For the Overarticulated condition, speakers were asked to "exaggerate" their articulation and to "speak very clearly, as if for someone with a hearing loss". For the Underarticulated condition, the instructions were to "mumble". If the contrasts desired were still unclear to the speakers, the experimenter demonstrated the three conditions.

In addition, subjects were asked to watch the VU meter of the tape recorder and to make sure the deflection of its needle stayed within a narrow range around the 0 dB mark. In this way, the average intensity of each sample was kept approximately equal across conditions and sentences. Using a metronome and a stop watch, the speakers also practiced keeping their speaking rates approximately equal across conditions. Speakers found that monitoring intensity and timing across conditions was difficult, since their natural tendency was to increase the intensity and duration of utterances in order to achieve articulatory clarity.

Each sentence was produced two or three times consecutively, until the speaker was satisfied with at least one utterance under each condition. Each speaker was permitted to rehearse as much as desired before recording commenced, but, even so, most speakers reported that they found the task difficult. The labelling of samples through the rest of this paper as Underarticulated, Overarticulated or Normally Articulated should therefore be taken to refer to the speakers' intentions rather than to the condition actually achieved.

3.3 DESIGN OF THE LISTENING TEST

3.31 PREPARATION OF THE LISTENING TEST TAPE

First, the best one of the two or three tokens of each utterance was selected and isolated, based on absence of hesitations, misarticulations and timing irregularities. Even so, the quality of the tokens varied across speakers and sentences, mainly due to unsuccessful rate control. Generally, utterances intended to be underarticulated were shortest, while those intended to be overarticulated were longest.

Because of the variable quality of the tokens, and because the length of the listening test needed to be limited so that the listeners could maintain their concentration, a subset of the best recorded tokens was selected for the Listening Test and the MI computations. The basis of selection was similarity of utterance duration across the three conditions for each speaker, while retaining as many phonemes as possible in the sentence material. One speaker, however, was excluded, because she was unable to produce contrasts of articulatory clarity to her own or to the experimenters' satisfaction. Also, one speaker (Speaker 1) was included in spite of the variability of his utterances, because his productions represented extremes in timing differences, and it was desirable to discover the effect of this variability on the MI values and on the listener judgments.

The sentences selected are listed in Appendix B, and information about the speakers is given in Table I. Eventually, three sentences each from the French speaker (Subject 6) and from six English speakers (Subjects 0 to 5) were selected, for a total of 63 tokens (3 sentences x 3 conditions x 7 speakers). See Table II for listings of utterance durations and ranges of durations for the samples selected.

Speaker	Sex	Language	Dialect Area
S0	F	English	Western Canadian
S1	M	English	Received Pronunciation
S2	F	English	Western Canadian
S3	F	English	Newfoundland
S4	M	English	Western Canadian
S5	M	English	Western Canadian
S6	M	French	Lausanne, Switzerland

Table I. Information regarding speakers selected.

		Articulatory Condition			
Speaker	Sentence	U	N	O	
S0	1	2.6 sec.	2.6	2.8	(0.2)
	2	2.4	2.5	2.9	(0.5)
	3	2.5	2.7	2.6	(0.2)
S1	1	1.6	2.1	2.8	(1.2)
	2	1.8	2.9	3.7	(1.9)
	3	2.0	2.5	3.5	(1.5)
S2	1	2.1	2.2	2.5	(0.4)
	2	2.1	2.5	2.6	(0.5)
	3	2.5	2.6	2.8	(0.3)
S3	1	2.3	2.3	2.2	(0.1)
	2	2.1	2.4	2.5	(0.4)
	3	2.3	2.5	2.5	(0.2)
S4	1	2.2	2.3	2.4	(0.2)
	2	2.3	2.3	2.6	(0.3)
	3	2.5	2.3	2.4	(0.2)
S5	1	2.7	3.0	2.5	(0.5)
	2	2.6	2.5	2.9	(0.4)
	3	2.9	2.4	2.7	(0.5)
S6*	1	1.9	1.9	2.0	(0.1)
	2	1.9	1.9	2.0	(0.1)
	3	1.9	2.0	2.0	(0.1)

* The durations listed for Speaker 6 are for corresponding French sentences.

Table II. Duration in seconds of the utterances selected. In parentheses, duration differences (in seconds) between shortest and longest token, for each set of three.

Once the selection process was completed, each selected token was copied twice onto the listening test tape in pseudo-random order, each item and its duplicate being separated by at least one other item. The beginning of each utterance was separated from the beginning of the next utterance by nine seconds. Since each utterance was approximately 2.5 seconds in length, this resulted in about 6.5 seconds of silence between successive utterances sample, an amount which was found to be satisfactory in a pilot test.

In order that the recording order could be checked, the speech samples were recorded on Channel 1 of the two track tape, and identification numbers for each test item were recorded on Channel 2. When the tape was played to the listeners, only the speech samples on Channel 1 were heard, but by setting the recorder to play both Channel 1 and 2 simultaneously, each item could be heard together with its identification number if the experimenters wished to identify any sample on the tape.

To summarize, the listening test tape consisted of 126 test items (3 sentences x 3 conditions x 7 speakers x 2 tokens of each utterance), plus ten practice items at the beginning and four dummy items (with response spaces on the answer sheet but no items recorded) at the end. The running time of the complete tape was 20.4 minutes.

3.32 LISTENERS

Ten English speaking listeners - five male and five female - were used for the Listening Test. No French listeners were included. The hearing of all listeners was tested beforehand using standard audiometric procedures. One male candidate listener was found to have a previously undetected hearing loss, and was thus replaced in the study. Two of the listeners had no knowledge of French, but the remaining eight had some knowledge, ranging from elementary knowledge to good fluency. Table III provides information regarding the listeners used in the perceptual test.

3.33 PROCEDURES FOR THE LISTENING TEST

The test tape was presented over Sennheiser HD 420 headphones. Listeners were asked to rate each utterance on a seven point scale. An example of the response sheet is given in Appendix C, as well as the instructions. The low (left) end of the scale was labeled "Underarticulated", the mid point "Normal", and the high (right) end "Overarticulated". After some practice items had been provided, the listeners had the opportunity to stop the tape recorder and ask questions about these items, or any part of the test, if they wished. Following this, the tape was rewound to the beginning, and the listeners were encouraged to continue through the entire test without stopping, if possible. The "dummy" items provided at the end of the tape were aimed at avoiding

Subject	Sex	Age (years)	Knowledge of French
L1	M	26	NONE
L2	F	25	SOME
L3	F	36	SOME
L4	F	26	SOME
L5	F	25	SOME
L6	F	33	SOME
L7	M	27	SOME
L8	M	29	SOME
L9	M	26	NONE
L10	M	22	SOME

Table III. Information regarding the listeners.

any end effects, such as rushing through in anticipation of finishing. They consisted of items numbered on the answer sheet which were not actually presented on the tape.

3.4 THE MODULATION INDEX

3.41 DESCRIPTION OF THE INDEX

A program was developed to compute a measure of amplitude modulation depth in the digitized envelopes of the speech samples. The program is listed in Appendix D.

Figure 7 illustrates the typical peaks and troughs found in a speech signal envelope. The program identifies first the peaks (a's) and troughs (b's) of the waveform

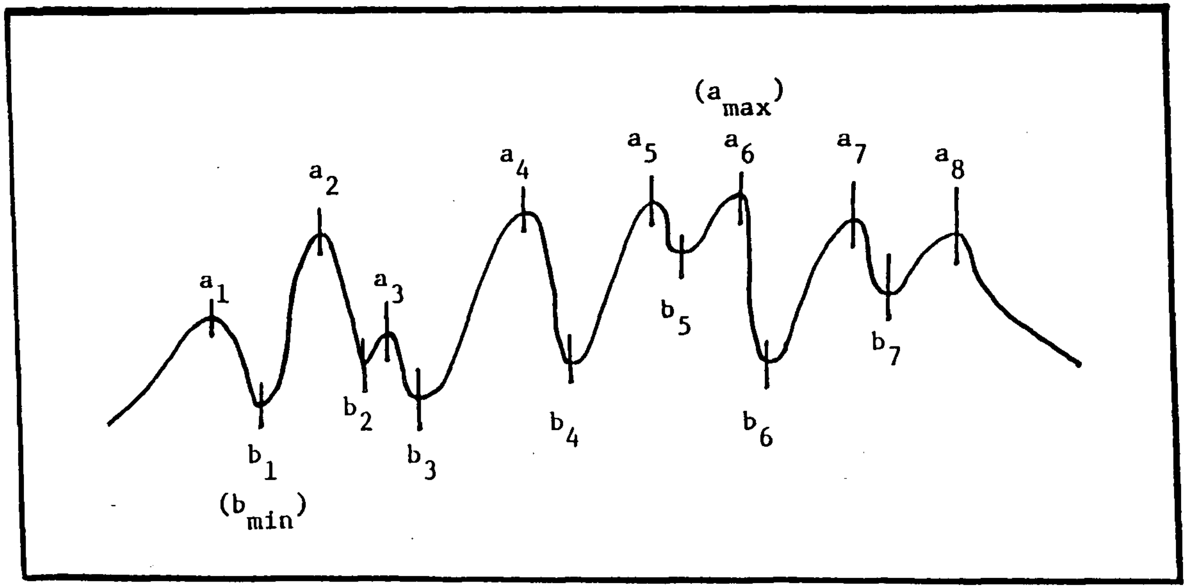


Figure 7. Labelling of peaks and troughs in the amplitude envelope of a speech sample.

envelope. The amplitude and the location of each peak/trough is then stored. In addition, the highest peak (a_{max}) and lowest trough (b_{min}) are determined, as well as their average (av = $(a_{max} + b_{min})/2$).

The ratios of trough-to-adjacent-peak amplitudes are calculated. The product of the obtained values is taken to be the basic measure of amplitude modulation in the sample, since as amplitude modulation depth increases (i.e. greater articulatory clarity), trough-to-peak ratio decreases and therefore the MI decreases. The geometric average of this product is taken to normalize MI values for tokens of different lengths. The basic formula for the calculation of of modulation depth is thus:

$$MI^2 = \sqrt[n]{\frac{av}{a_1} \cdot \frac{b_1}{a_1} \cdot \frac{b_1}{a_2} \cdot \frac{b_2}{a_2} \cdot \frac{b_2}{a_3} \cdots \frac{b_{n-1}}{a_n} \cdot \frac{b_n}{a_n}} \ ,$$

or more simply

$$MI^2 = \sqrt[n]{\frac{av}{a_1} \cdot \frac{b_1^2}{a_2^2} \cdot \frac{b_2^2}{a_3^2} \cdots \frac{b_{n-2}^2}{a_{n-1}^2} \frac{b_{n-1}^2}{a_n^2}} \ .$$

The squaring of terms is then eliminated by taking the square root of both sides of the equation to yield finally the MI:

$$MI = \sqrt[n]{\frac{av}{a_1} \cdot \frac{b_1}{a_2} \cdot \frac{b_2}{a_3} \cdots \frac{b_{n-2}}{a_{n-1}} \cdot \frac{b_{n-1}}{a_n}} \ ,$$

where n is the number of peaks, av is a value equal to $(a_{max} + b_{min})/2$, the a 's are the peak values, and the b's are the trough values.

3.42 DIGITIZATION OF TOKENS FOR MI CALCULATIONS

The use of various versions of the digitized envelope signal was investigated. The basic method is illustrated in Figure 8. After rectification, the signal was low-pass filtered with cutoff frequencies of either 25 Hz or 75 Hz. In some cases, this smoothing was followed by logarithmic amplification; in others, it was followed by linear amplification. Eventually, the method resulting in the least smoothing of the envelope was chosen - i.e. 75 Hz low-pass filtering followed by linear amplification. It was reasoned that if smoothing was kept to a minimum, loss of amplitude modulation in the digitized envelopes would be avoided. The signal was sampled on a PDP-12 computer at 200 Hz and stored on LINC tape, using a set of programs developed by Lloyd Rice at UCLA.

Once stored, the signals could be displayed on the oscilloscope screen, in wave (graphic) form or in numerical form. Each signal was inspected individually, and starting and end points for MI computation were chosen. The starting point chosen was always on the rising slope of the first peak of the utterance, and the end point chosen was always on the falling slope of the last peak (see Figure 9).

Trough amplitudes with negative values had to be avoided because the geometric averaging implicit in the the MI formula cannot deal with them. Similarly, a trough amplitude of zero is undesirable since it would result in a calculated MI value of zero. For these reasons, each digitized envelope was manipulated through a program so that it contained no digitized trough amplitudes which were

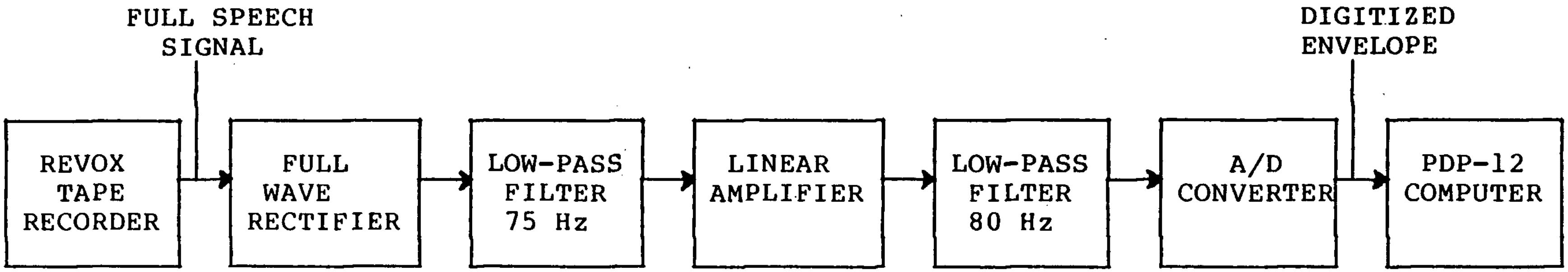


Figure 8. Speech sample digitization scheme.

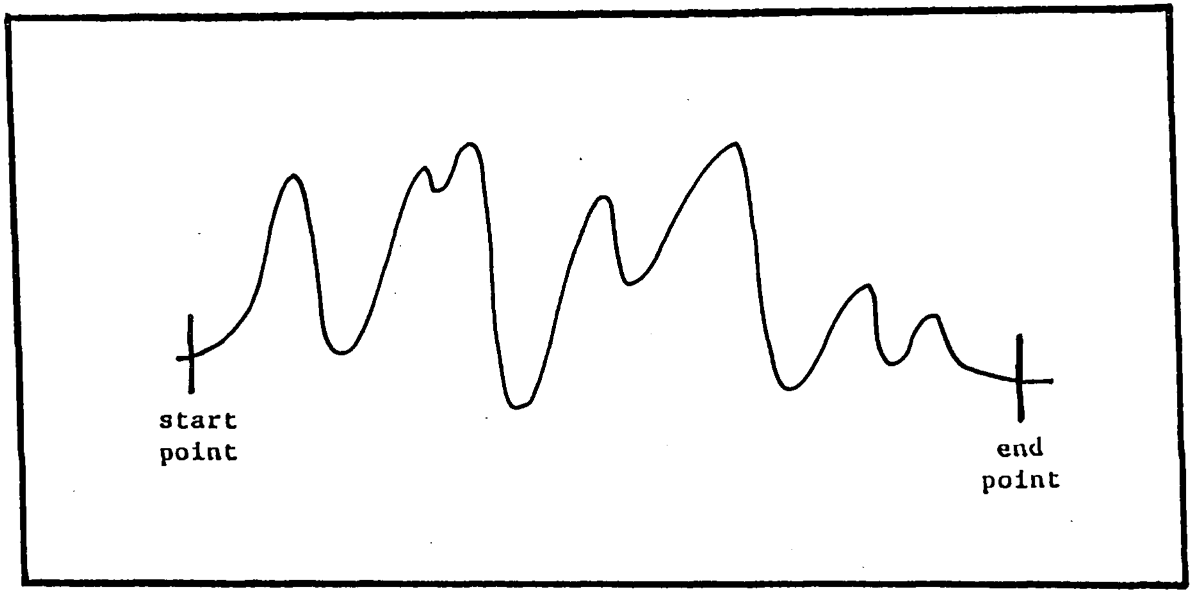


Figure 9. An example of start and end point locations chosen for Modulation Index analysis of the amplitude envelope of a speech sample.

negative or equal to zero. At the same time, since the intensities of the samples were found to vary somewhat, despite efforts to ensure uniformity through appropriate instructions to the speakers, each digitized envelope was adjusted upward or downward in such a way that the average amplitudes of all its peaks had approximately the same value for all utterances.

MI values were obtained and analyzed only for those utterances selected for the listening test.

CHAPTER FOUR

RESULTS

4.1 RESULTS OF THE LISTENING TEST

4.11 CONSISTENCY OF LISTENER JUDGMENTS

Listeners were asked to judge the articulatory clarity of the speech samples. They gave their judgments on an integer scale, from 1 (Mumbled) to 7 (Overarticulated). The answer sheet given to the listeners is illustrated in Appendix C. The correlation between the M'I values computed and the perceptual data could therefore be checked. The perceptual data also provided a check of how well the speakers performed, since their intentions were not necessarily realized in every case.

Each speaker's performance was evaluated by analyzing how the listeners judged his/her utterances. Each listener made 18 judgments about a given speaker's productions since each speaker produced a total of 9 tokens (3 sentences x 3 conditions), and each token was presented twice to the listeners. These 18 judgments were grouped, according to the three conditions intended by the speakers, into three sets of six judgments each.

The standard deviations of listener judgments for utterances produced by Speakers 0 to 6 are shown in Table IV as a function of the articulatory condition intended by the speaker - i.e. Underarticulated (U), Normally Articulated (N), and Overarticulated (O). Ten listeners made two judgments for each of three sentences per speaker, for a total number of sixty judgments per speaker and per condition. The listener judgments of speech samples from Speakers 2 and 6 were the least variable, and those of Speaker 5 were the most variable. These results indicate that listeners found utterances produced by Speaker 5 more difficult to judge than the utterances of other speakers. For this reason, the data from Speaker 5 was excluded from further analyses.

A similar analysis of standard deviations, this time as a function of the listener who made the judgments, revealed that judgments by Listener 6 were considerably less consistent than judgments by the other listeners. Table V shows the standard deviations of judgments for each listener across the articulatory conditions intended by the speakers (Speaker 5 excluded).

In addition, the performance of the listeners themselves was evaluated by analyzing the consistency of their judgments for repeated items (hereafter "repeatability"). The results of the analyses of repeatability are shown in Table VI as a function of the speaker whose utterances were judged, and in Table VII as a function of the listener who made the judgments. Further exclusions of speakers or listeners were not necessary on the basis of these results.

Speaker	Under- articulated s.d. (n=60)	Normal articulation s.d. (n=60)	Over- articulated s.d. (n=60)	Mean s.d. (combined conditions) (n=180)
0	0.58	0.86	0.72	0.72
1	0.83	0.70	0.60	0.71
2	0.41	0.72	0.59	0.57
3	0.56	0.76	0.91	0.74
4	0.67	0.61	1.02	0.77
5	0.73	0.87	0.96	0.85
6	0.53	0.51	0.64	0.56

Table IV. Standard deviations for listener judgments across the articulatory conditions intended by the speakers for Speakers 0 to 6 inclusive. (The data were drawn from 10 listeners, and two judgments per listener per sentence.)

Listener	Under- articulated s.d. (n=42)	Normal articulation s.d. (n=42)	Over- articulated s.d. (n=42)	Mean s.d. (combined conditions) (n=126)
1	0.54	0.59	0.64	0.59
2	0.47	0.84	0.92	0.74
3	0.81	0.60	0.63	0.68
4	0.58	0.68	0.63	0.63
5	0.44	0.23	0.54	0.40
6	0.79	1.18	1.18	1.05
7	0.61	0.73	0.90	0.75
8	0.75	0.67	0.81	0.74
9	0.63	0.90	0.67	0.73
10	0.57	0.73	0.85	0.72

Table V. Standard deviations for listener judgments by ten listeners across the articulatory conditions intended by the speakers. (The data are drawn from Speakers 0 to 6 inclusive, and from two judgments per sentence for three sentences per speaker.)

Speaker	Number of differences greater than 1 between repeated judgments	Percent differences greater than 1 between repeated judgments (n=81)
0	9	11.1
1	8	9.9
2	4	4.9
3	5	6.2
4	9	11.1
6	11	13.6

Table VI. Repeatability of listener judgments as a function of the speaker. (The data are drawn from three articulatory conditions, three sentences per speaker, and nine listeners - Listener 6 excluded.)

Listener	Number of differences greater than 1 between repeated judgments	Percent differences greater than 1 between repeated judgments (n=54)
1	1	1.9
2	8	14.8
3	5	9.3
4	6	11.1
5	1	1.9
7	3	5.6
8	6	11.1
9	10	18.5
10	10	18.5

Table VII. Repeatability of listener judgments across listeners. (The data are drawn from three articulatory conditions, and three sentences for each of six speakers - Speakers 0, 1, 2, 3, 4, and 6.)

For all other analyses of the perceptual data, the mean of the two judgment scores from each of the listeners for each token replaced the individual scores.

4.12 COMPARISON OF LISTENER'S JUDGMENTS WITH SPEAKER'S INTENTIONS

In Figure 10, the means of listener judgment scores, separately for each speaker, are plotted as a function of the articulatory condition intended by the speaker. The same data are arranged in Figure 11 to display the means of listener judgment scores, all speakers pooled, for each listener separately. The means of the judgment scores in each condition can be seen to correlate well with the speakers' intentions for every speaker. When individual sentences were considered, however, some disagreement between the speakers' intentions and the listeners' perception became apparent. In addition, the agreement between speaker's intentions and listener's perception was better for some speakers and some listeners than for others. Speakers 2 and 6, and Listeners 1, 3, and 5 were the best subjects in this respect.

For the three-way contrast of articulatory conditions (U vs. N vs. O), there were 162 sets of judgments, since 6 speakers each produced 3 sentences which were each judged by 9 listeners. The listeners' perception agreed with the

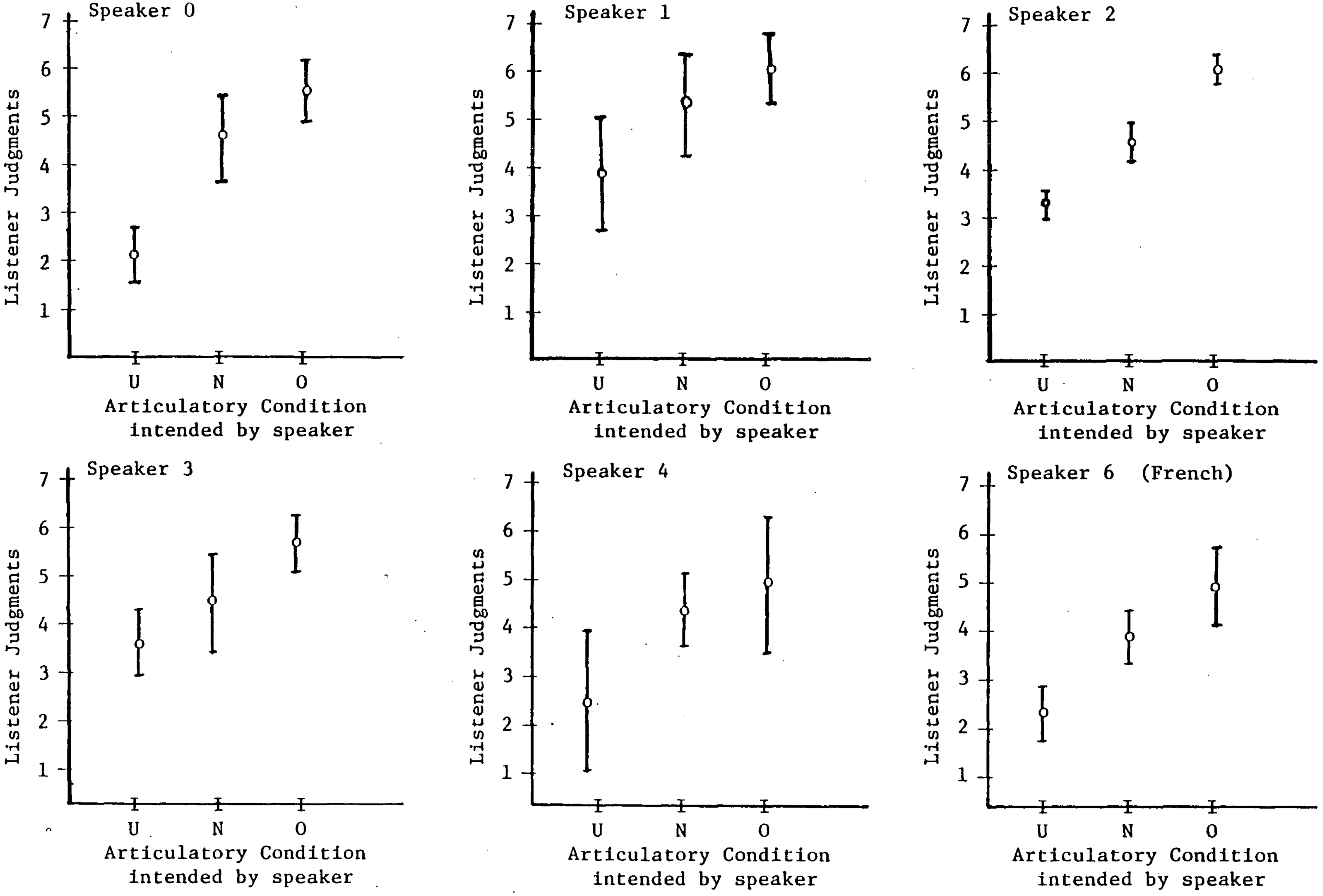


Figure 10. Listener judgments as a function of the speaker. o = mean judgment. I = two standard deviations.

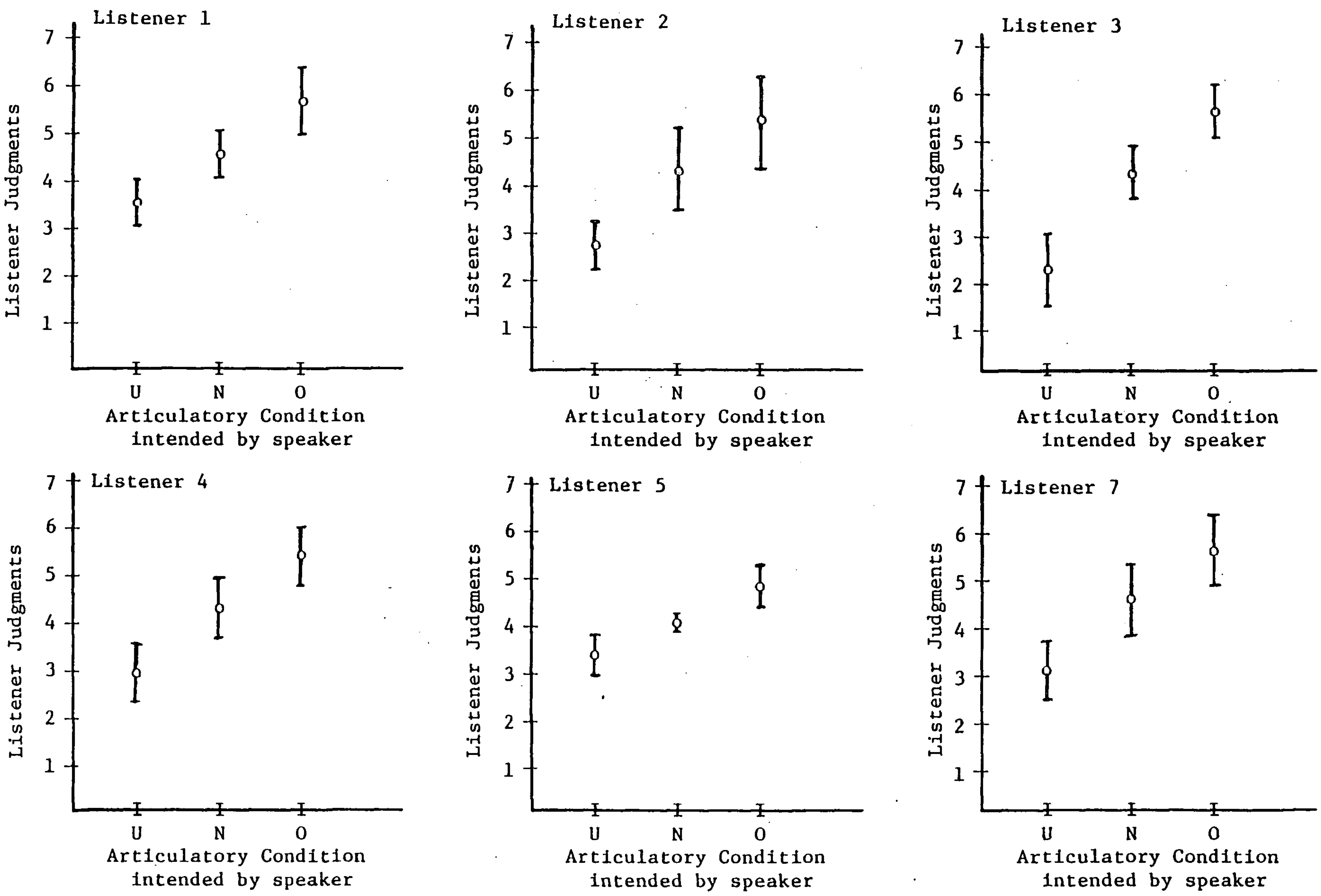


Figure 11. Listener judgmen-s as a function of the listener. o = mean judgment. I = two standard deviations. (continued...)

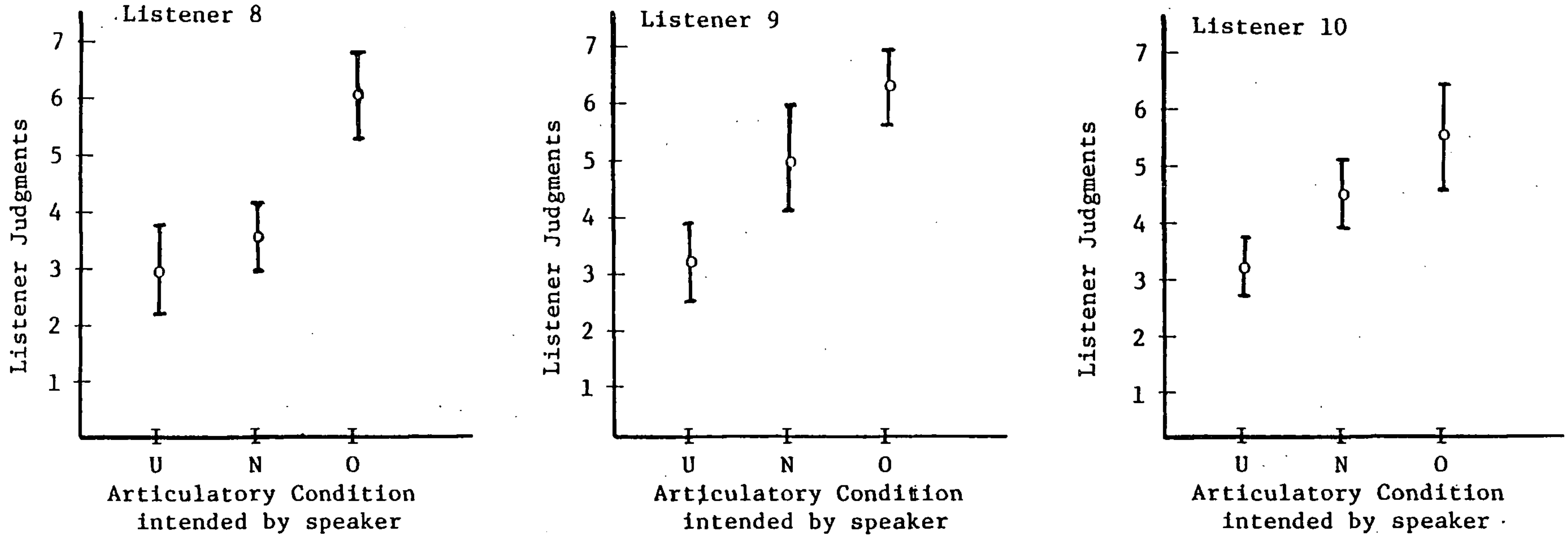


Figure 11. (...continued)

speaker's intentions in 117 out of 162 sets. Agreement for the two-way contrasts was 88% (143/162) for U vs N, 83% (135/162) for N vs. O, and 99% (161/162) for U vs. O. Thus, there was, in general, a monotonic relationship between the speakers' intentions and the listeners' perception of increasing articulatory clarity.

4.13 EFFECT OF LANGUAGE ON LISTENER JUDGMENTS

The small size of the corpus used limited the conclusions which could be drawn about the effects of language on the data. However, based on the data available, agreement between the speakers intentions and the perceptual data was slightly better for the English than the French speech. Of the 27 sets of listener judgments applying to the three French sentences, (9 listeners x 3 sentences) 17 (63%) had complete agreement between the speaker's intentions and the relative ranking of articulatory clarity assigned by the listeners for all three articulatory conditions. For the English speakers, there was 73% (99/135 judgments) agreement between speakers' intentions and listeners' judgments.

The familiarity of the listeners with French also seems to affect their judgments. In fact, when only the French utterances are considered, the two listeners who knew no French were in 100% (6/6 sentences) agreement with the speaker's intentions, whereas the 7 listeners with some knowledge of French agreed with the speaker's intentions for only 11 of 21 sentences (52%). Perhaps the listeners with no

French were able to concentrate more on the articulation of the speaker than the listeners with some French, since the latter group may have been distracted by attempts to decipher the semantic content of the sentences. The small sample size does not allow statistical analysis of the difference observed or pursuit of this possible explanation, however.

4.14 EFFECT OF UTTERANCE DURATION VARIABILITY ON LISTENER JUDGMENTS

Speaker 1 was selected for inclusion in the Listening Test because he, of all the speakers, produced the greatest contrasts of utterance duration across the three articulatory conditions. However, a t-test shows that speaker/listener agreement for Speaker 1 (22/27 judgments or 81%) was not significantly different from the agreement for the other English speakers (77/108 judgments or 71%) ($p = 0.05$, $z = 1.55$ on a two-tailed Proportions Test).

4.2 RESULTS OF THE MI CALCULATIONS

Various combinations of low-pass filtering (25 Hz, 75 Hz) and amplification (logarithmic vs. linear) were explored in digitizing the data in view of MI computation. The parameters selected for the final measurements were low-pass filtering with a cutoff frequency of 75 Hz, and linear amplification. Of all the variations explored, this combination resulted in retention of the greatest amount of amplitude modulation in the digitized sample. It was reasoned that setting the cutoff frequency of the filter as high as possible (given equipment limitations) and using linear rather than logarithmic amplification should avoid any loss of contrast in amounts of amplitude modulation for different samples which might result otherwise from excessive smoothing of the envelope.

The MI values obtained with this analysis are listed in Table VIII as a function of the speaker's intentions for each utterance. The MI values were arranged into MI ranking orders (see Table IX). A smaller MI value indicates more amplitude modulation in the signal. Consequently, in accordance with the hypothesis that amplitude modulation increases with articulatory clarity, MI ranking is assigned as a function of decreasing MI value. The order of increasing articulatory clarity intended by the speaker was assumed to correspond to the order UNO for each sentence. Each of the tokens retained its label according to speaker intentions, but the order of the three tokens per sentence could be rearranged according to MI values. For example, a sentence with MI values of 0.789 for the U condition, 0.876 for the N condition, and 0.688 for the O would be assigned an MI ranking of NUO.

As can be seen in Table IX, the MI ranking order failed

		Modulation Index Values		
Speaker	Sentence	Under-articulated	Normal articulation	Over-articulated
0	1	0.819	0.906	0.923
	2	0.800	0.884	0.926
	3	0.936	0.865	0.832
1	1	0.886	0.848	0.897
	2	0.860	0.872	0.868
	3	0.871	0.847	0.902
2	1	0.767	0.586	0.683
	2	0.885	0.813	0.842
	3	0.779	0.808	0.828
3	1	0.894	0.762	0.860
	2	0.896	0.885	0.909
	3	0.860	0.844	0.788
4	1	0.821	0.674	0.699
	2	0.889	0.822	0.833
	3	0.831	0.795	0.791
6	1	0.842	0.875	0.778
	2	0.923	0.916	0.752
	3	0.839	0.868	0.751

Table VIII. Modulation Index values for utterances by six speakers. (The division of values by articulatory condition reflects the intentions of the speakers.)

Speaker	Sentence	Modulation Index ranking
0	1	ONU
	2	ONU
	3	UNO
1	1	OUN
	2	NOU
	3	OUN
2	1	UON
	2	UON
	3	ONU
3	1	UON
	2	OUN
	3	UNO
4	1	UON
	2	UON
	3	UNO
6	1	NUO
	2	UNO
	3	NUO

Table IX. Modulation Index rankings of utterances by articulatory condition. (The labels for each token corresponds to the articulatory condition intended by the speaker. According to the experimental hypothesis, all MI ranking orders would be UNO if the speakers were completely successful at producing the desired articulatory contrasts.)

to match the speakers' intentions consistently. Agreement between MI ranking and speakers' intentions was 22% (4/18 sentences) when a three-way contrast (U vs. N vs. O) was considered. For the three two-way contrasts, U vs. N., N vs. O, and U vs. O, agreement was 72% (13/18), 39% (7/18), and 61% (11/18), respectively.

Sample sizes were too small to allow meaningful analysis of the effects of language or of utterance duration variabiltiy on MI values.

4.3 COMPARISON OF MI VALUES WITH PERCEPTUAL DATA

4.31 COMPARISON OF MI VALUES WITH LISTENER JUDGMENTS

MI values ranking orders were compared to the listener judgments ranking orders. Each token still retained its label according to the speaker's intentions. There were 18 possible comparisons, one for each of three sentences produced by each of the six speakers. However, some way of selecting the sentences for which there was most agreement between listeners was needed. The only sentences out of the original eighteen included in the analysis were those for which all nine listeners had previously shown agreement with the speakers' intentions, or sentences for which one listener out of nine had judged two tokens to have the same degree of articulatory clarity. In other words, at worst, one listener had judged as equivalent two different tokens of the same sentence.

Five sentences (1 each produced by Speakers 0 and 1, and 3 produced by Speaker 2) met the above criteria in all three conditions. This meant that, in the three-way contrast U vs. N vs. O, a perceptual ranking order of UNO could be assigned on the basis of the judgments from at least eight listeners. None of the five sentences had an MI ranking order of UNO (0% agreement). For the two-way contrasts U vs. N, N vs. O, and U vs. O, agreement was 71% (10/14 sentences), 27% (3/11 sentences), and 61% (11/18 sentences), respectively. Both the perceptual data and the MI values indicate that the speakers produced more contrast in articulatory clarity between the U and N conditions than between the N and O conditions.

In order to better evaluate the relationship between the MI values and the perceptual data, data from the best speakers (Speaker 2 and Speaker 6) was plotted against data from the best listeners (Listeners 1, 3, and 5) (see Figure 12). Unexpectedly, trends were suggestive of a non-monotonic relationship with lower MI values for tokens judged by listeners to be over- or underarticulated than for tokens judged to be normally articulated. There was, however, a great deal of scatter in these plots, indicating that the relationship between the MI and this perceptual data is not particularly strong.

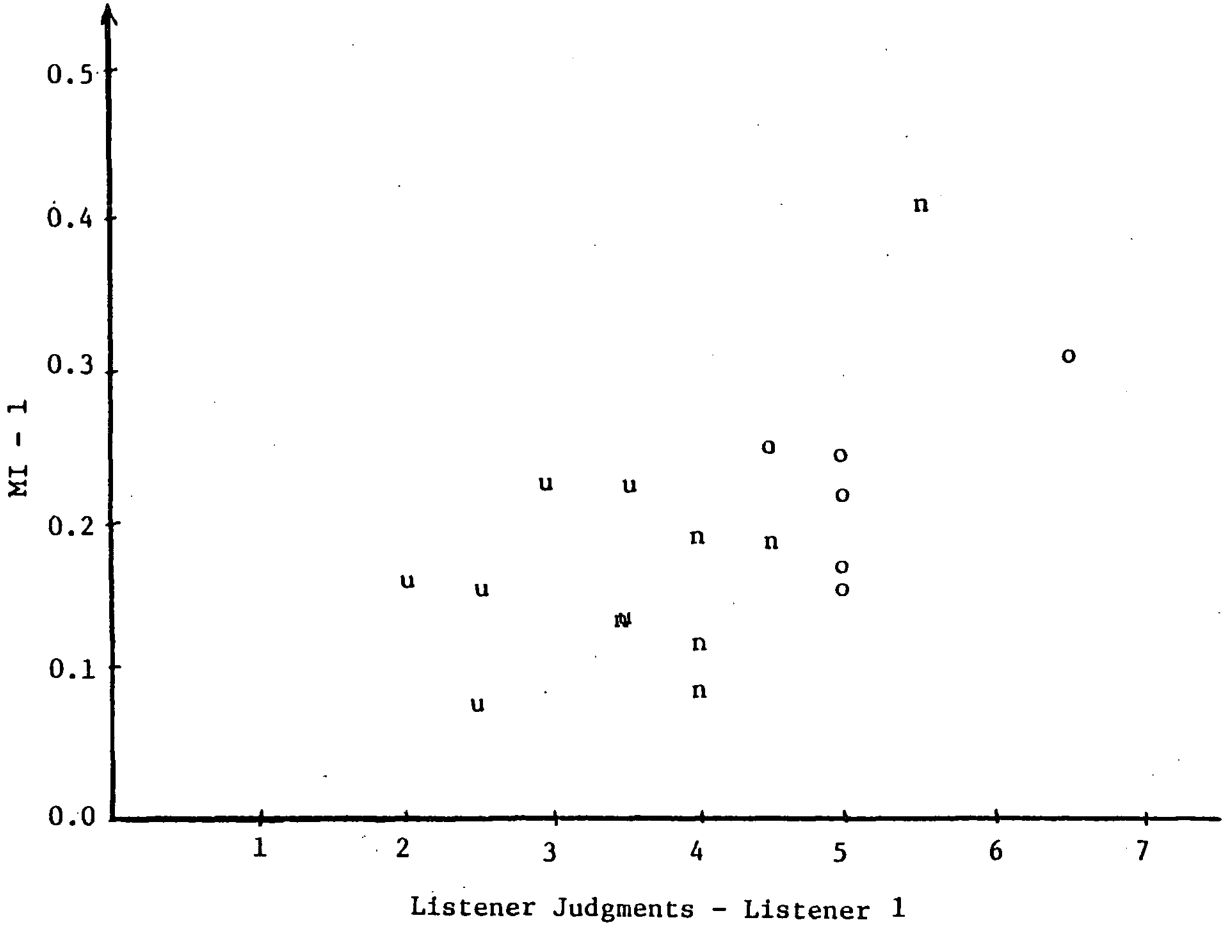


Figure 12a. The relationship between Modulation Index values and listener judgments. MI values are for utterances by Speakers 2 and 6. u indicates the point corresponds to an utterance intended by the speaker to be Underarticulated. n indicates the point corresponds to an utterance intended by the speaker to be Normally Articulated. o indicates the point corresponds to an utterance intended by the speaker to be Overarticulated.

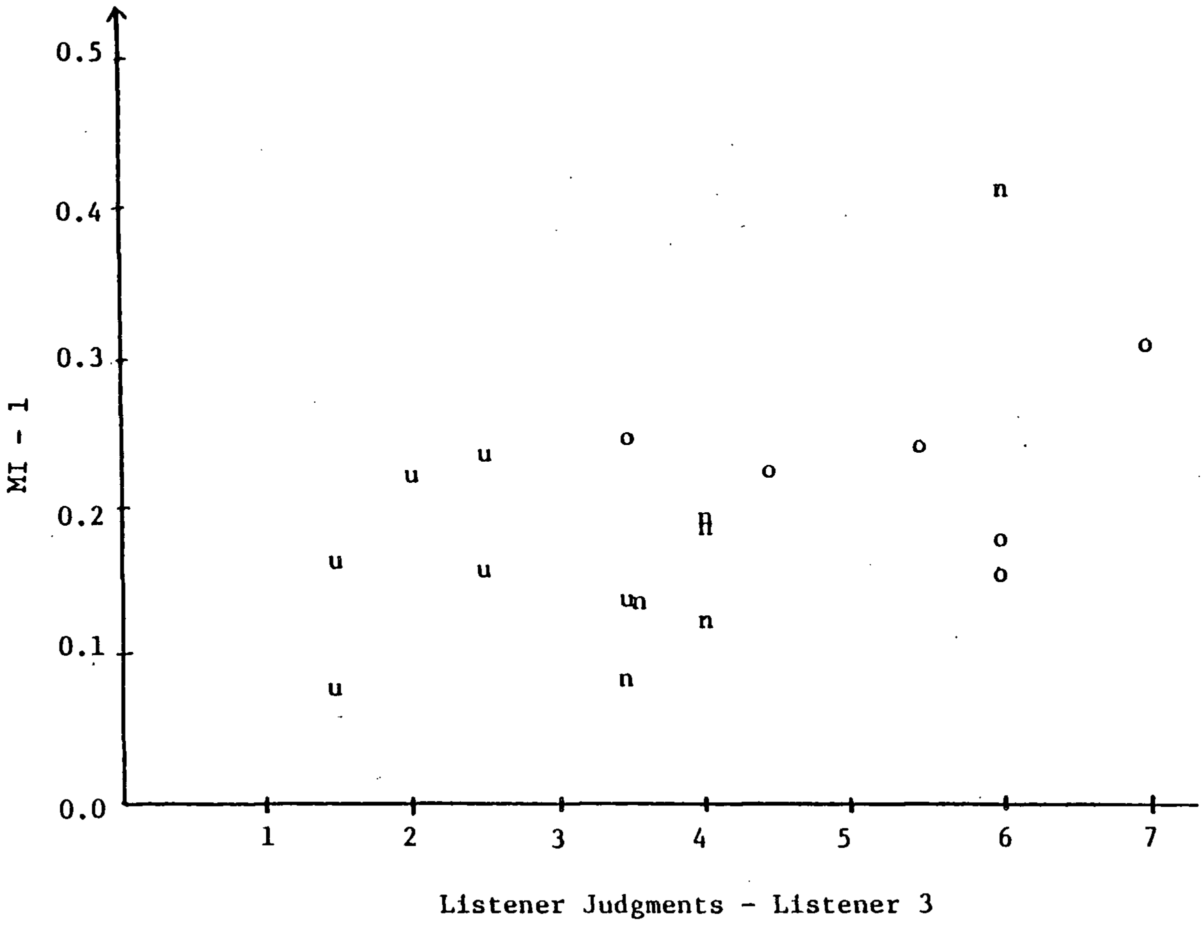

0.5
0.4
0.3
0.2
0.1
0.0
MI - 1
1
2
3
4
5
6
7
Listener Judgments - Listener 3
u
u
u
u
u
u
u
n
n
n
n
n
n
o
o
o
o
o
o
o

Figure 12 b.

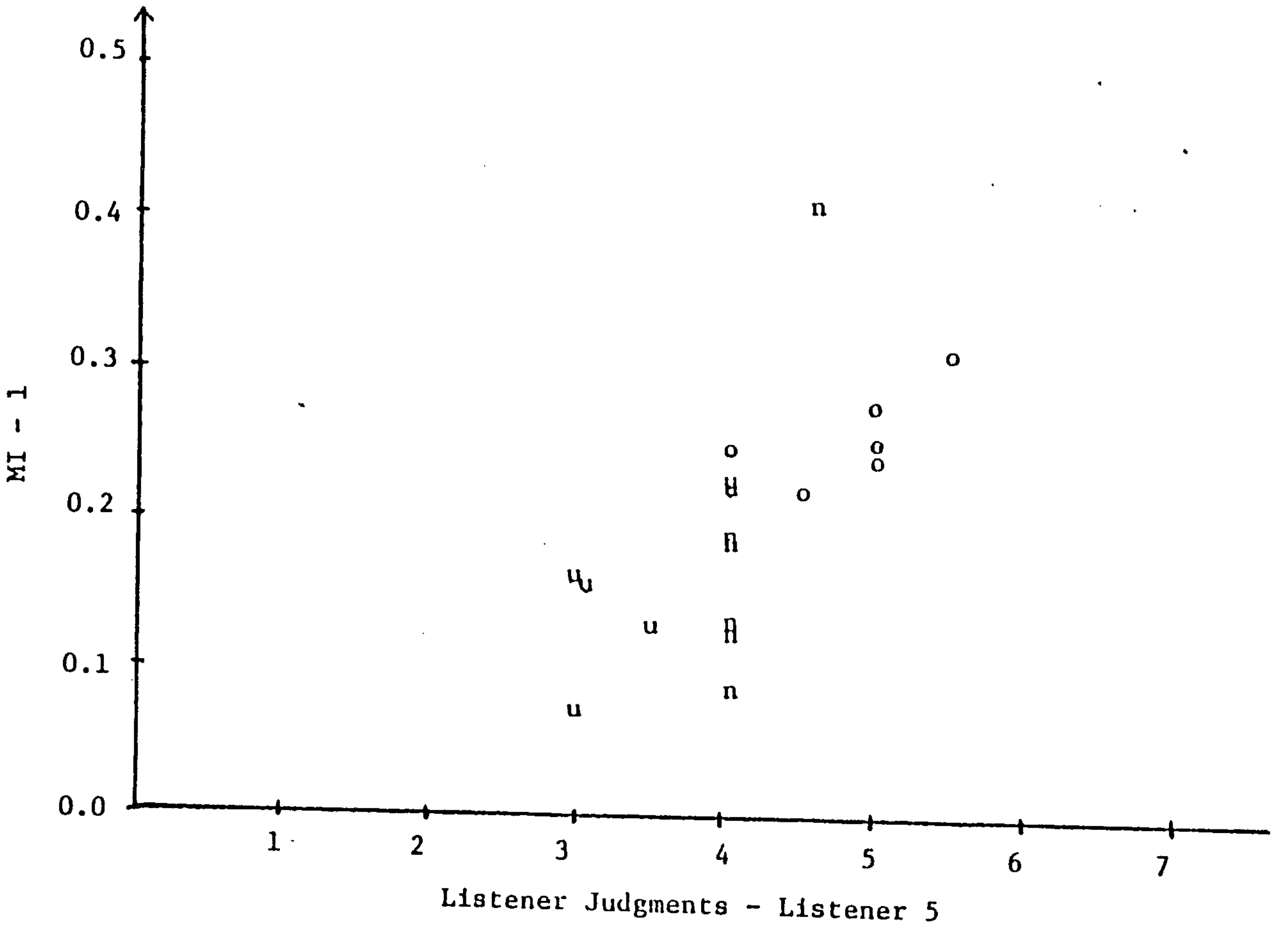

0.5
0.4
0.3
0.2
0.1
0.0
MI - 1
1
2
3
4
5
6
7
Listener Judgments - Listener 5
n
o
o
o
o
o
o
o
u
n
n
u
u
u
n
n
u

Figure 12c.

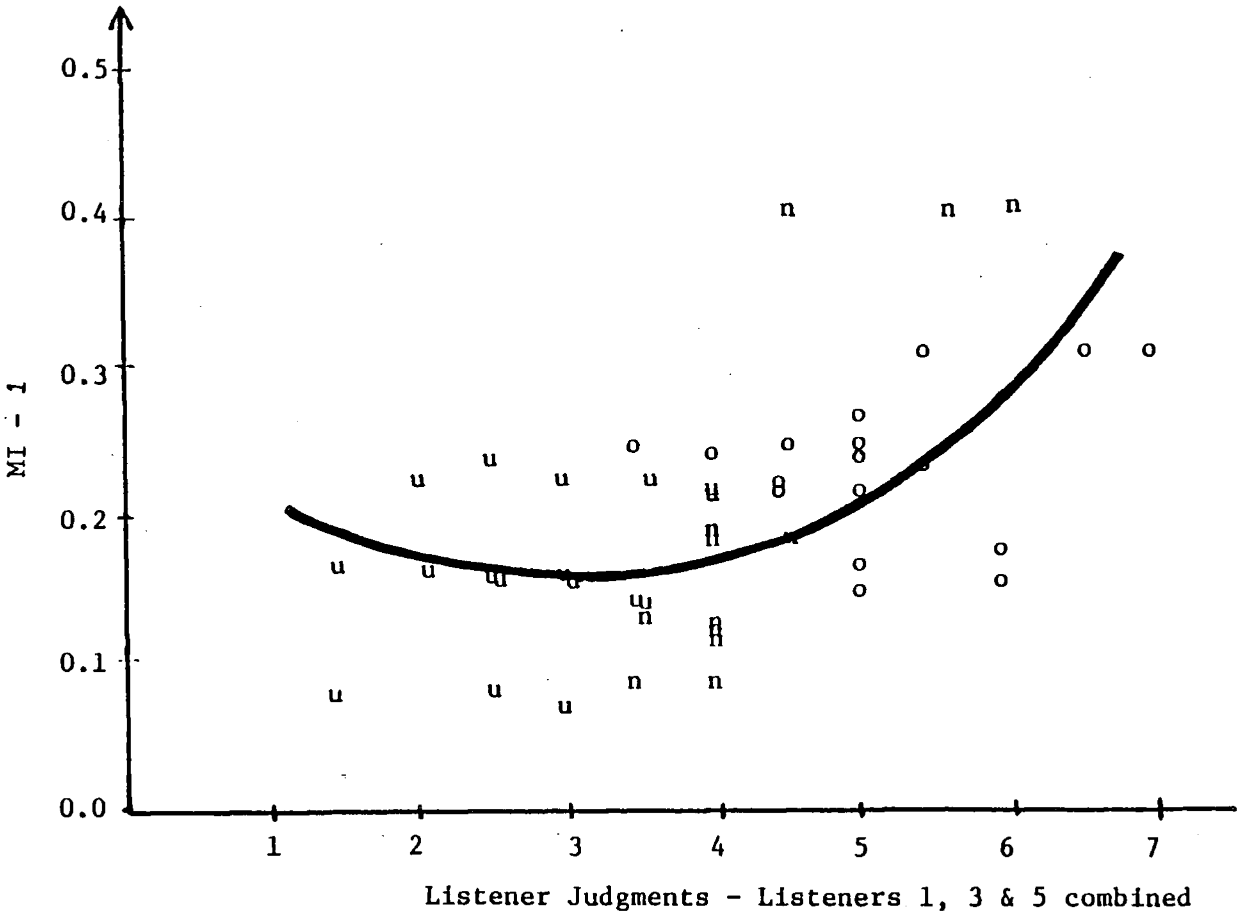


Figure 12d.

The line indicates the non-monotonic relationship suggested by the points.

4.32 COMPARISON OF MI VALUES WITH VISUAL IDENTIFICATION OF WAVEFORMS

In an informal blind test, the author attempted to visually identify the intended articulatory conditions (U, N, or O) from the waveforms of the various randomly selected samples, when they were displayed on an oscilloscope after amplitude equalization. The visual identifications matched the speaker intentions in only 11 out of 30 trials. A confusion matrix (Table X) reveals that the author could readily discern the U vs. O contrast intended by the speakers, but not the U vs. N vs. O contrast. The visual identifications were also compared to the perceptual ranking orders based on the listeners' judgments. For this comparison, a confusion matrix (Table XI) again showed that tokens which were judged to have low or high articulatory clarity by the listeners were those visually identified by the author as U or O respectively, whereas tokens which were rated in the middle of the scale of articulatory clarity by the listeners were visually identified as O most often, as U some of the time, but seldom as N.

Three differences were observed in the shapes of the waveforms from the three articulatory conditions. First, Overarticulated tokens, as hypothesized, presented the most depth of modulation, which could be clearly observed by comparing the excursion of adjacent peaks and troughs.

		Articulatory Condition visually identified		
		U	N	O
Articulatory Condition intended by the Speakers	U	5	5	0
	N	4	1	5
	O	1	4	5

Table X. Confusion matrix comparing visual identification of articulatory condition with speaker intentions.

		Articulatory Condition visually identified		
		U	N	O
Articulatory Condition based on rankings of tokens according to Listener Judgments	U	55	47	6
	N	16	8	30
	O	10	34	62

Table XI. Confusion matrix comparing visual identification of articulatory condition with the articulatory condition suggested by the perceptual ranking order according to the listeners' judgments.

Second, whereas Underarticulated tokens appeared smooth and rounded, with a few main peaks and troughs, Overarticulated tokens tended to have jagged peaks, and the top of each main peak contained several smaller peaks and troughs. The Normally Articulated tokens had features of the other two conditions. The third observable difference deserves more discussion.

The Overarticulated tokens, and to a lesser extent, the Normally Articulated tokens, had plateaus in their envelopes where the intensity dropped to zero for some short period of time. These plateaus corresponded to pauses between words. This resulted in a "bottom-clipping" effect, since the downward excursion of troughs was limited to zero intensity. As can be seen from the formula for the calculation of the MI, the MI was not designed to take bottom clipping into account. Recall that:

$$MI = \sqrt[n]{\frac{av}{a_1} \cdot \frac{b_1}{a_2} \cdot \frac{b_2}{a_3} \cdot \frac{b_3}{a_4} \cdots \frac{b_{n-1}}{a_n}} \quad ,$$

where n is the number of peaks, av = $(a_{max} + b_{min})/2$, a_{max} is the value of the highest peak, and b_{min} is the value of the lowest trough, $a_{1 \text{ to } n}$ are the peak values, and $b_{1 \text{ to } n-1}$ are the trough values.

Samples were manipulated prior to MI calculations so that zero trough amplitudes would be eliminated. Therefore, since MI values depend on trough-to-peak amplitude ratios, if trough amplitudes are limited by bottom clipping when peak amplitudes are not, MI values will be artificially elevated for tokens containing pauses. This elevation of MI values would apply mainly to tokens which are more clearly articulated, and for which low MI values are expected according to the hypothesis which is the motivation for this study. It is possible, then, that this elevation of MI values weakened the apparent relationship between the MI ranking order and the perceptual ranking order by reducing contrasts between MI values for tokens judged by the listeners to differ in articulatory clarity.

This possibility was explored with a subset of the data. It was reasoned that MI values would be less affected by limited excursion of troughs if the MI formula were arithmetic instead of geometric.

The MI formula was modified to include differences between peak and trough values rather than ratios. The following is the modified MI formula:

$$\mathrm{MImod} = 1/n[av - a_n + \sum_{i=1}^{n-1}(b_i - a_i)]$$

$$av = 1/2(a_{max} + b_{min}).$$

MI was then recalculated for a subset of tokens which contrasted in the number of pauses they contained. The resulting MI ranking orders are shown in Table XII.

The modification does not change the MI ranking orders significantly. For the two sentences for which there is a change in ranking (Speaker 3, Sentence 1 and Speaker 4, Sentence 3), the separation in the original MI values

Speaker	Sentence	MI ranking	MImod ranking
0	1	ONU	ONU
2	2	UON	UON
3	1	UON	UON
4	3	UNO	UON
6	1	NUO	NUO

Table XII. Comparison of rank orders according to the original MI values with rankings according to MImod. Peak/trough ratios are the basis for MI calculations; peak/trough differences are the basis for MImod scores.

between the two conditions which reversed was very small, and consequently the changes in rank cannot be considered significant.

There was an improvement in most cases in the separation between MI values calculated for tokens of contrasting articulatory clarity with this modification. Even so, the lack of improvement in rank ordering of conditions suggests that some other method of dealing with bottom clipping, such as elimination of pauses from the tokens analyzed, may be a better solution.

CHAPTER FIVE

DISCUSSION AND CONCLUSIONS

The purpose of this study was the exploration of a method for estimating the intelligibility of different speakers as affected by the articulatory clarity of their speech. The basic premise behind this computed measure, the MI, is that the amplitude envelope of the waveform for intelligible speech is characterized by greater amplitude modulation than the amplitude envelope for less intelligible speech; consequently, a measure of amplitude modulation should provide a measure of speech intelligibility. This idea has been previously applied to the more general case of speech intelligibility in listening spaces by Houtgast, Steeneken, and their colleagues (Houtgast and Steeneken, 1973, 1984; Houtgast, Steeneken and Plomp, 1980; Plomp, Houtgast and Steeneken, 1980; Steeneken and Houtgast, 1980; van Reitschote, Houtgast and Steeneken, 1981, 1984; Steeneken and Houtgast 1983). These authors developed the Modulation Transfer Function as the basis for the second version of the Speech Transmission Index.

In room acoustics, the changes between source and receiver of the amplitude spectrum of a speech sample may be compared before and after passage through a speech transmission system. Reduction in amplitude modulation and the corresponding reduction in intelligibility are due to several causes, in particular noise, reverberation, and the filtering effect of the system.

In the second version of the STI, Houtgast and Steeneken (Houtgast and Steeneken, 1973, 1985; Steeneken and Houtgast, 1980, 1983) used a simple artificial signal consisting of a sum of sinusoidally modulated bands of pink noise as an input signal for measurement of the Modulation Transfer Function. The signal is thus shaped across the frequency range of 125 to 8000 Hz to resemble an average speech amplitude spectrum modulated at the modulation frequencies found in natural speech. In contrast, when comparing intelligibility of one speaker, natural speech must be used, since it is not the average properties of speech which are of interest. This makes the task of the proposed measure, the MI, somewhat more complicated than the task of the Modulation Transfer Function. The modulation frequencies in natural speech are unlikely to remain constant across speakers and across speech samples, and a measure using natural speech must also be equipped to cope with intensity and timing differences.

In order to minimize the effects of these other factors in the present study, a design was used in which speakers attempted to produce the same sentences in three different ways: Underarticulated (or mumbled), Normally Articulated, and Overarticulated. Efforts were made to minimize differences in intensity and duration of tokens, through instruction to the listeners, selection of the tokens with least variability in duration across articulatory conditions, and manipulation of samples to equalize intensities. Even so, contaminants such as intensity and timing variations within the tokens remained. One source of error in this study, therefore remained: the speakers, more specifically the type of speech they produced. Although speakers were allowed rehearsal time and feedback as to the adequacy of their productions, the task of producing the articulatory contrasts proved to be difficult, and some speakers were more successful than others in producing articulatory contrasts.

This problem was compounded when the MI values were checked against perceptual data. Normal hearing listeners were asked to rate the speakers' utterances as to the articulatory clarity they perceived. Results indicate that the listeners were fairly successful in this respect - much more so, in fact, than the MI. However, some listeners were much better (and more consistent) than others in guessing the speaker's intentions. Perhaps some of the listeners were responding to additional cues in the speech tokens (i.e. timing or intensity differences, etc.) which the MI was not designed to detect. The observation that listeners with no knowledge of French seemed better at judging the articulatory clarity intended by a French speaker than listeners with no knowledge of French suggests that the listeners could be distracted by factors other than those of interest to the experimenters.

In spite of all of the difficulties encountered, a relationship was observed between the MI values (calculated for the utterances of the best speakers) and the judgments of the best listeners (see Figure 12). The trend was toward a non-monotonic relationship, with lower MI values for tokens judged to be Underarticulated or Overarticulated than for those tokens judged to be Normally Articulated. The nature of this relationship presents a problem. According to the non-monotonic curve suggested by the distribution of the points, low MI values correlate with both extremely high and extremely low articulatory clarity, whereas high MI values correlate with average articulatory clarity. Clearly, some modifications to the measure are required in order to establish a monotonic relationship between MI values and articulatory clarity, if the MI is to be a useful speech intelligibility rating tool.

In order to discover what some of these modifications should be, the experimenter studied the envelopes of the tokens displayed on an oscilloscope. There were cues which served to identify the articulatory conditions. As expected, the modulation depth observed in the tokens increased with the articulatory clarity perceived by the listeners. In addition to this expected contrast, however, the number of pauses between words increased as perceived articulatory clarity increased. Although according to Picheny, Durlach and Braida (1985, 1986) the number of pauses is a factor contributing to speech intelligibility, for the purposes of MI calculations of amplitude modulation, the pauses were contaminants. Since pauses are brief periods of silence, the intensity at pauses descends to zero and then remains there until the next word begins. This results in a "bottom clipping" effect which is not accounted for in the calculation of the MI. A modification to the MI formula designed to reduce the effect of bottom clipping on MI scores was tested without satisfactory results. A better solution may be the selection of speech tokens which do not contain pauses for MI analysis, or addition of a correction factor to MI values obtained from speech tokens containing pauses.

A final modification to the MI which we were unable to explore, due to limitations in equipment and time, was frequency dependent analysis. French and Steinberg (1947) were the first to employ frequency dependent analysis when calculating the Articulation Index. They divided the speech spectrum into twenty frequency bands, each of which was calculated to contribute equally to intelligibility in the ideal case. A more modern approach has been to employ octave bands or third-octave bands, and to compare the signal within each of the bands selected before and after passage through a transmission system. The advantage of this division of the signal is that frequency-specific effects, such as low-pass filtering, interfering narrow band noise, or frequency-specific amplitude modulation, can be measured with more precision. These effects may be important to the intelligibility of the signal, but their significance may be lost in wide band analysis which averages the frequency band(s) of interest with the other unaffected bands.

To summarize, the concept of rating speech intelligibility of speakers by quantification of amplitude modulation in speech amplitude envelopes seems promising, based on the results of this exploratory study. However, the MI needs modification before it becomes a useful tool. In particular, the effects of inter-speaker variations in timing and intensity need to be overcome adequately. Although some suggestions for modifications to the MI have been presented here, further research will be necessary to discover the form of the Index which will be most effective.

BIBLIOGRAPHY

Ahlstrom, C. and Humes, L.E. (1983). "Normal modulation transfer functions in noise using temporal probe tones," ASHA, 25(10):67.

Ahlstrom, C. and Humes, L.E. (1984). "Psychoacoustic modulation transfer functions of impaired ears using probe tones," ASHA, 26(10):175.

Ahlstrom, C., Boney, S.F. and Humes, L.E. (1984). "Modulation transfer functions of hearing aids," ASHA 26(10):124.

Ananthapadmanabha, T.V. (1982) "Intelligibility carried by speech source functions: implications for a theory of speech perception," Speech Transmission Laboratory, Quarterly Report. STL-QPSR 4/1982.

Ando, K. and Canter, G.J. (1969). "A study of syllabic stress in some English words by deaf and normally hearing speakers," Lang.Speech 12:247-255.

ANSI (1969). American National Standard methods for the calculation of the articulation index (ANSI S3.5 - 1969) New York: ANSI.

Black, J.W. (1957). "Multiple choice intelligibility tests," J. of Speech and Hearing Dis. 22(2):213-240.

Brannon, J.B. (1966). "The speech production and spoken language of the deaf. Lang.Speech 12:247-255.

Crum, M.A. (1974). "Effects of speaker to listener distance upon speech intelligibility in reverberation and noise," Doctoral dissertion. Northwestern University.

Dareham, J.R. (1986). "Measuring speech intelligibility using RASTI," Sound Video Contractor 3(11)50,52,54,56-57.

Dirks, D.D., Bell, T.S., Rossman, R.N. and Kincaid, G.E. (1986). "Articulation index predictions of contextually independent words," J.Acoust.Soc.of America 80:82-92.

Doyle, J. (1987). "Reliability of audiologists' ratings of the intelligibility of hearing impaired children's speech," Ear and Hearing 8(3):170-174.

Dunn, H.K. and White, S.D. (1940). "Statistical measurements on conversational speech," J.Acoust.Soc.of America 11:278-288.

Fant, G. (1960). Acoustic Theory of Speech Production Mouton, Hague. Second Edition 1970.

French, N.R. and Steinberg, J.C. (1947). "Factors governing the intelligibility of speech sounds," J.Acoust.Soc.of America 19:90-119.

Hirsh, I.J., Reynolds, E.G. and Joseph, M. (1954). "Intelligibility of different speech materials," J.Acoust.Soc.of America 26:530-538.

Houtgast, T. and Steeneken, H.J.M. (1971). "Evaluation of speech transmission channels by using artificial signals," Acustica 25:355-367.

Houtgast, T. and Steeneken, H.J.M. (1973). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica 28:66-73.

Houtgast, T. and Steeneken, H.J.M. (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J.Acoust.Soc.of America 77:1069-1077.

Houtgast, T. and Steeneken, H.M.J. (1984). "A multi-language evaluation of the RASTI - method for estimating speech intelligibility in auditoria," Acustica 54:185-199.

Houtgast, T., Steeneken, H.J.M. and Plomp, R. (1980). "Predicting speech intelligibility in rooms from the modulation transfer function I. General room acoustics," Acustica 46:60-72.

Hudgins, C.V. (1960). "The development of communication skills among profoundly deaf children in an auditory training programme," In I.R. Ewing (Ed.), Modern Educational Treatment of Deafness. Manchester (Eng.): Manchester University Press.

Hudgins, C.V., and Numbers, M. (1942). "An investigation of the intelligibility of the speech of the deaf," Genet.Psycholog. Monogr. 25:289-392.

Humes, L.E., Dirks, D.D., Bell, T.S., Ahlstrom, C. and Kincaid, G.E. (1986). "Application of the Articulation Index and the Speech Transmission Index to the recognition of speech by normal-hearing and hearing impaired listeners," J. of Speech and Hearing Res. 29:447-462.

John, J.E.J., and Howarth, J.N., (1965). "The effect of time distortions on the intelligibility of deaf children's speech," Lang.Speech 8:127-134.

Kamm, C.A., Dirks, D.D. and Bell, T.S. (1985). "Speech recognition and the articulation index for normal and hearing-impaired listeners," J.Acoust.Soc.of America 77:281-288.

Klein, W. (1971). "Articulation loss of consonants as a basis for the design and judgment of sound reinforcement systems," J.Audio Eng. soc. 19:11:920.

Kondraske, G.V. (1985). "Quantitative assessment of speech function," Proc. 7th. Annual Conference of the IEEE Medical Biological Society. Chicago, Illinois. September 27-30, 1985. Volume 2:671-674.

Kryter, K.D. (1962a). "Methods for the calculation and use of the articulation index," J.Acoust.Soc.of America 34:1689-1697.

Kryter, K.D. (1962b). "Validation of the articulation index," J.Acoust.Soc.of America 34:1698-1702.

Kryter, K.D. and Ball, J.H. (1964). "SCIM - A meter for measuring the performance of speech communication systems," Decision Science Laboratory. Electronic systems Division. Air Force Systems Command. Report ESD-TDR-64-674.

Licklider, J.C.R., Bisberg, A. and Schwarzlander, H. (1959). "An electronic device to measure the intelligibility of speech," Proceedings of the National Electronics Conference 15:329.

Lochner, J. and Burger, J. (1964). "The influence of reflections on auditorium acoustics," J.Sound Vibr. 4,426-454.

Lundin, F.J. (1982). "The influence of room reverberation on speech - an acoustical study of speech in a room," Speech Transmission Laboratory, Quarterly Report. STL-QPSR 4/1982.

Lundin, F.J. (1986). "A study of speech intelligibility over a public address system," Speech Transmission Laboratory, Quarterly Report. STL-QPSR 1/1986.

Metz, D.E., Sama R., V.J., Schiavetti, N., Sitler, R. and Whitehead, R.L. (1985). "Acoustic dimensions of hearing impaired speaker's intelligibility," J. of Speech and Hearing Res. 28:345-355.

Miller, G.A. and Nicely, P.E. (1955). "An analysis of perceptual confusions among some English consonants," J.Acoust.Soc.of America 27:338-352.

Monsen, R.B. (1978). "Toward measuring how well deaf children speak," J. of Speech and Hearing Res. 21:197-219.

Morse, P.M. and Ingard, U.K. (1968). Theoretical Acoustics McGraw-Hill. N.Y.

Newman, P.W (1979). "Appraisal of articulation," Part II in Darley, F.L. (Editor) Evaluation of Appraisal Techniques in Speech and Language Pathology. Don Mills, Ontario: Addison-Wesley Publishing Company.

Pavlovic, C.V. (1987). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J.Acoust.Soc.of America 82:413-422.

Pavlovic, C.V. and Studebaker, G.A. (1984) "An evaluation of some assumptions underlying the articulation index," J.Acoust.Soc.of America 75:1606-1612.

Peutz, V.M.A. (1971). "Articulation loss of consonants as a criterion for speech transmission in a room," J.Audio.Eng. 19(11):915.

Picheny, M.A., Durlach, N.I. and Braida, L.D. (1985). "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," J. of Speech and Hearing Res. 28:96-103.

Picheny, M.A., Durlach, N.I. and Braida, L.D. (1986). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. of Speech and Hearing Res. 29:434-446.

Plomp, R., Houtgast, T. and Steeneken, H.J.M. (1980). "Predicting speech intelligibility in rooms from the modulation transfer function II. Mirror image computer model applied to rectangular rooms," J.Acoust.Soc.of America 46:73.

Rettinger, M.A. (1968). Acoustics: Room Design and Noise Control New York: Chemical Publishing Co., Inc.

Schroeder, M.R. (1981). "Modulation transfer functions," Acustica 49:179.

Steeneken, H.J.M. and Agterhuis, E. (1978). "Description of STIDAS II-c (Speech Transmission Index Device using Artificial Signals) Part I," Report 1978-19,IZF-TNO, Soesterberg, The Netherlands.

Steeneken, H.J.M. and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality," J.Acoust.Soc.of America 67:318-326.

Steeneken, H.J.M. and Houtgast, T. (1983). "The temporal envelope spectrum of speech and its significance in room acoustics," In Proceedings of the Eleventh International Congress on Acoustics. 7:85-88.

Steeneken, H.J.M. and Houtgast, T. (1985). "RASTI: A tool for evaluating auditoria," Bruel & Kjaer Tech. Rev. 3.

van Reitschote, H.F., Houtgast, T. and Steeneken, H.J.M. (1981). "Predicting speech intelligibility in rooms from the modulation transfer function IV: A ray-tracing computer model," Acustica 49:245-252.

van Reitschote, H.F., Houtgast, T. and Steeneken, H.J.M. (1984). "Predicting speech intelligibility in rooms from the modulation transfer function V: The merits of a ray-tracing model versus general room acoustics," Acustica 53:72-78.

Voelker, C.H. (1938) "An experimental study of the comparative rate of utterance of deaf and normal hearing speakers," Am.Ann.Deaf 83:274-284.

Williams, C.E. and Hecker, M.H.L. (1968). "Relation between intelligibility scores for types of speëch distortion. J.Acoust.Soc.of America 44:1002.

APPENDIX A

INTELLIGIBILITY TESTS BASED ON LISTENER JUDGMENTS

A. TESTS CITED BY BLACK (1957)

Bell Telephone's Tests
reference: Fletcher, H. and Steinberg, J.C. (1929). "Articulation testing methods," Bell Syst. Tech. J.8:806-854.

Harvard's Phonetically Balanced (PB) Tests
references: a) Egan, J.P. (1944). "Articulation testing methods II," Psycho-Acoustic Laboratory, Harvard University, Nov. OSRD Rept. no. 3820. b) Egan, J.P. (1948). "Articulation testing methods," Laryngoscope 58:955-991.

Voice Communication Laboratory's Test
reference: Haagen, C.H. (1944). "Intelligibility measurement: Techniques and procedures used by the Voice Communication Laboratory," Psychological Corporation., New York, OSRD Rept. no. 3748.

B. TESTS REVIEWED IN NEWMAN'S (1979) CHAPTER

Drumwright, A.F. (1971). The Denver Articulation Screening Examination (DASE) Ladoca Project and Publishing Foundation, Inc.

Fisher, H.A. and Logemann, J.A. (1971). The Fisher-Logemann Test of Articulation Competence Houghton Mifflin Company.

Fudala, J.B. (1963). Arizona Articulation Proficiency Scale Western Psychological Services. (later editions 1970, 1974).

Goldman, R. and Fristoe, M. (1969). Goldman-Fristoe Test of Articulation (GFTA) American Guidance Service, Inc. (later edition published 1972).

Hejna, R.F. (1968). Developmental Articulation Test Speech Materials.

Irwin, O.C. (1972) "Integrated articulation test," In Orvis C. Irwin, Communication Variables of Cerebral Palsied and Mentally Retarded Children Springfield, Ill: Charles C. Thomas.

McDonald, E.T. (1964). A Deep Test of Articulation, Picture Form Stanwix House, Inc. (Also available from the same publisher are the Sentence and Screening Forms of the same test).

Mecham, J.L.J. and Jones, J.D. (1970). Screening Speech Articulation Test (SSAT) Communication Research Associates, Inc..

Pendergast, K., Dickey, S., Selmar, J.W., and Soder, A.L. (1969). Photo Articulation Test (PAT) Interstate Printers and Publishers, Inc..

Templin, M.C. and Darley, F.L. (1969). The Templin-Darley Tests of Articulation University of Iowa Bureau of Educational Research and Service, Sound edition.

Toronto, A.S. (1977). Southwestern Spanish Articulation Test (SSAT) Academic Tests, Inc..

Van Riper, C. and Erickson, R.L. (1973). Predictive Screening Test of Articulation (PSTA) Western Michigan University, Continuing Education Office, Third Edition.

C. TESTS CITED BY WILLIAMS AND HECKER (1968)

Harvard PB-Word Intelligibility Test and
Harvard Sentence Tests
reference: Egan, J.P. (1948) "Articulation testing methods," Laryngoscope 58:955-991.

Fairbanks Rhyme Test
reference: Fairbanks, G. (1958) "Test of phonemic differentiation: The Rhyme Test," J.Acoust. Soc.Amer. 30:596-600.

Modified Rhyme Test
reference: House, A.S., Williams, C.E., Hecker, M.H.L and Kryter, K.D. (1965) "Articulation testing methods: Consonantal differentiation with a closed-response set," J.Acoust. Soc. Amer. 37:158-166.

APPENDIX B

THE ENGLISH AND FRENCH SENTENCES SELECTED

ENGLISH

1. Patty put five pennies in her purse.
2. Shallow seas are not shark infested.
3. Green grapevines grow in country gardens.

FRENCH

1. Il n'y a jamais de fumée sans feu.
2. Il joue du trombone tous les lundis.
3. Un grand coca-cola sans glaçons.

APPENDIX C

LISTENING TEST INSTRUCTIONS AND ANSWER SHEET

PART 1 - LISTENING TEST ANSWER SHEET

Subject ID: L

In this experiment, you will be hearing a number of sentences spoken by six English speakers and one French speaker. Sometimes the sentences are mumbled (underarticulated), sometimes they are articulated normally, and sometimes they are overarticulated. Your task is to assign a number on a seven point scale for each sentence indicating how it sounds to you. For example, if you were fairly certain that a sentence was normally articulated, you would circle the number 4 as shown below:

mumbled normal overarticulated

141, 1 2 3 4 5 6 7,

whereas if you thought that item 142 was mumbled, you might circle 1 as shown below:

mumbled normal overarticulated

142. 1 2 3 4 5 6 7

The scale represents a continuum between mumbled (number 1) and overarticulated (number 7). You may choose any number which you think is appropriate to what you hear.

Please try to attend only to whether the sentence sounds mumbled, normally articulated, or overarticulated, and ignore other variables such as differences in recording or voice quality, speed of articulation, language spoken, or sentence content.

Please begin by listening to the first ten sentences <u>without</u> marking the paper, and then stop the tape and ask questions if you need clarification about any part of the task. Following this, the tape will be rewound to the beginning and you will be asked to listen to the entire tape and mark your answer sheet, without rewinding or stopping the tape if possible.

Subject ID: L

	mumbled			normal		overarticulated	
1.	1	2	3	4	5	6	7
2.	1	2	3	4	5	6	7
3.	1	2	3	4	5	6	7
4.	1	2	3	4	5	6	7
5.	1	2	3	4	5	6	7
6.	1	2	3	4	5	6	7
7.	1	2	3	4	5	6	7
8.	1	2	3	(...etc...)			
129.	1	2	3	4	5	6	7
130.	1	2	3	4	5	6	7
131.	1	2	3	4	5	6	7
132.	1	2	3	4	5	6	7
133.	1	2	3	4	5	6	7
134.	1	2	3	4	5	6	7
135.	1	2	3	4	5	6	7
136.	1	2	3	4	5	6	7
137.	1	2	3	4	5	6	7
138.	1	2	3	4	5	6	7
139.	1	2	3	4	5	6	7
140.	1	2	3	4	5	6	7
	mumbled			normal		overarticulated	

APPENDIX D

LISTING OF THE FOCAL-12 PROGRAM FOR CALCULATING THE MODULATION INDEX

PART A

```
1.01  C PROGRAM DHPP12A
1.03  E
1.05  O C
1.07  L O,F0,I,INPUT,1
1.08  L O,F0,I,OUTPUT,0
1.10  A "FIRST SAMPLE",SF,!,"LAST SAMPLE",SL,!;S I=SF+2;S
K=1
1.20  D 2;G 4.1

2.10  S A=F0(I)-F0(I-1);S B(0)=F0(I)-F0(I+1)
2.20  S B(I)+F0(I+1)-F0(I+2);S B(2)=F0(I+2)-F0(I+3);S
B(3)=F0(I+3)-F0(I+4);S B(4)=F0(I+4)-F0(I+5)

4.10  I (A)4.2,4.3;I (B(0))5.4,5.2,5.1
4.20  I (B(0))5.6,5.3,5.4
4.30  I (B(0))5.2,5.5,5.2

5.09  C IT IS A PEAK
5.10  S F1(K)=F0(I);S F1(K+1)=I;S K=K+2;G 6.1;C STORING
AMPLITUDE AND LOCATION OF PROUGHS
5.19  C SHOULD NOT HAPPEN
5.20  G 7.1
5.29  C LOOK ONE, TWO, OR THREE AHEAD
5.30  I (A*B(1))5.8,5.31;I (A)5.34,7.1,5.34
5.31  I (A*B(2))5.85,5.32;LI (A)5.35,7.1,5.35
5.32  I (A*B(3))5.9,5.33;I (A)5.37,7.1,5.37
5.33  I (A*B(4))5.95,7.1;I (A)5.38,7.1,5.38
5.34  S F1(K)=F0(I);S F1(K+1)=I+0.5;S I=I+1;S K+K+2;G 6.1
5.35  S F1(K)=F0(I);S F1(K+1)=I+1.O;S I=I+2;S K=K+2;G 6.1
5.37  S F1(K)=F0(I);S F1(K+1)=I+1.5;S I=I+3;S K=K+2;G 6.1
5.38  S F1(K)=F0(I);S F1(K+1)=I+2.0;S I=I+4;S K=K+2;G 6.1
5.39  C CONTINUE
5.40  G 6.1
5.49  C SHOULD NOT HAPPEN
5.50  G 7.1
5.59  C IT IS A TROUGH
5.60  S F1(K)=F0(I);S F1(K+1)=I;S K=K+2;G 6.1
5.80  S I=I+1;G 6.1
5.85  S I=I+2;G 6.1
5.90  S I=I+3;G 6.1
5.95  S I=I+4;G 6.1

6.10  S I=I+1;I (I-SL+5)1.2,1.2
6.20  T %2.01,!!,I,!,K,!!!,"TYPE RETURN TO CONTINUE";A !!
6.30  S F1(0)=(K-1)/2;C STORES IN F1(0) THE NUMBER OF
"PROUGHS"
```

```
6.40   L C,F1
6.50   L O,DHPP12B,0
```

PART B

```
1.01   C PROGRAM DHPP12B
1.02   E
1.03   O C
1.05   L O,F1,F,OUTPUT,0
1.10   S MAX=0;S MIN=1000;S K=F1(0)
1.20   F I+1,2,K*2;D 2
1.30   D 3
1.40   S PR=AV/F1(IS);F I=IS+2,4,IL-2;S PR=PR*F1(I)/F1(I+2)
1.50   S MI=FEXP((1/NP)*FLOG(PR))
1.60   T %7.06,"MI = ",MI,!!
1.90   Q

2.01   C DETERMINING MAX AND MIN
2.10   I (F1(I)-MAX)2.2,2.2;S MAX=F1(I)
2.20   I (MIN-F1(I))2.3,2.3;S MIN=F1(I)
2.30   S AV=(MA+MI)/2;R

3.01   C DETERMINING THE LOCATION OF THE FIRST AND LAST PEAKS
3.10   I (F1(I)-F1(3))3.2,3.9;S IS=1;G 3.3
3.20   S IS=3
3.30   I (F1(K*2-1)-F1(K*2-3))3.4,3.9;S IL=K*2-1;G 3.5
3.40   S IL=K*2-3
3.50   S NP=(IL-IS)/4+1;S NT=NP-1;R
3.90   T "TROUBLE',!;Q
```