QUANTIFYING THE DEGRADATION OF AUTOMATIC SPEECH RECOGNITION FOR REVERBERANT ENVIRONMENTS

By

Stephen Secules

A Thesis Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
Major Subject: ARCHITECTURAL SCIENCES

Approved:

____
Jonas Braasch, Thesis Adviser

____
Paul Calamia, Member

Rensselaer Polytechnic Institute
Troy, New York

July 2008
(For Graduation August 2008)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

Thank you to Luigi Rosa for providing the recognition program, and for fielding the several follow-up questions above and beyond.

Thank you to Jonas for putting up with constant check-in meetings and even more email flow. You've always put my interests first and tried to make it fun. Thanks to Paul for keeping me company while I waited for Jonas.

Thanks to my classmates for a weird and wonderful year. It was good to have someone in the bunkers with me. Thanks also to my family and friends who have supported me and helped me to this achievement.

ABSTRACT

In general, automatic speech recognition (ASR) algorithms are designed for use with pure, anechoic signals. Although ASR systems approach a human's intelligibility for clean signals, it is well acknowledged that a downfall of speech recognition systems is that recognition accuracy depreciates dramatically for signals with reverberation. Little is known about the specific character of this depreciation, its primary causes (e.g. room geometry, reverberation strength) or effects (blurring of syllables, plosives, consonants). The focus of this study is to precisely quantify the depreciation of speech recognition accuracy for reverberant signals using a black box experiment to vary reverberation characteristics and observe speech recognition accuracy. The methodology tests two speech recognition platforms on a standard recognition task for human speech perception, testing small vocabulary sets of similar sounding words. A range of reverberant settings was simulated by convolution with an impulse response. The artificial reverberant settings were developed using an image source model of simple geometries and small rooms, varying the average room absorption. The impulse response was also divided into various energy balances of early reflections and late energy, to determine the contribution to recognition depreciation from each component. The recognizers had the least reverberant recognition accuracy for words which only differed by their ending consonants. The depreciation of recognition accuracy from early reflections alone was lower than the overall room effect; however the overall depreciation with respect to the absorption coefficient was well predicted by the strength of the reverberant tail. The results were compared to the results of prior research. The reported results will help to characterize the problem for the automatic speech recognition community, and serve as a model for further precise investigation of the effects of room acoustics on developed algorithms.

1. INTRODUCTION

1.1 Motivation

Automatic speech recognition (ASR) is a technology designed to facilitate communication between a speaking human user and computer listener. The computer typically uses a statistical algorithm to convert measured acoustical data to a meaningful string of words. Though ASR technology has wide potential applications, it has been implemented in commercial products for a limited set of uses [1]. Its most prominent implementation is hands-free operation of personal computers and telemarketing communication. The potential stands for significant improvements in intelligence processing and closed captioning if an ASR implementation was more fully developed. The main barrier to an ASR implementation in these new applications is source quality. An ASR system is typically trained to respond in an ideal setting—noiseless and anechoic. When source material is not ideal due to noise or reverberation, the system's accuracy decreases dramatically, rendering it unreliable and useless. This is in stark contrast to human speech perception in reverberation, which depreciates much more slowly; this discrepancy points to a potential for improvement.

Improvement of ASR in cases of missing information due to noise or reverberation represents a significant joint research effort in the speech science and computer science communities. The speech community has long focused on the microscopic features of human speech perception and its breakdown in reverberation, (e.g. the importance of clarity in the 2-3 kHz frequency bands for hearing consonants). Contrarily, the speech recognition community is dominated by a computer-science based iterative design process, which seeks the best implementation to solve the macroscopic recognition problem by improving overall recognition accuracy. The speech recognition research is distinctly lacking in scientific understanding of the microscopic level reasons for and characteristics of the recognition breakdown. Thus, while the iterative design approach has led to a refinement of the current ASR algorithm methods to their maximized efficiency, it has not refined the definition of the problem in a parallel manner. A well-documented and scientific characterization of the problem will better inform the design process and point to the specific aspects of the algorithms which need improvement. There is also the possibility that the results could point to an entirely new approach to the recognition process, perhaps a process which is less of a statistical computerized measurement algorithm and closer to the current understanding of neurological processing of speech.

1.2 Research Focus

The focus of the current this research is to quantify the degradation of ASR algorithms in reverberant settings. The research attempts to add to the literature by combining an understanding of ASR systems and speech sciences with a basis and approach in the science of architectural acoustics. This research will add precision to the general conclusion that reverberation degrades recognition accuracy. It attempts to evaluate which specific phoneme properties are disrupted and by which specific room acoustics properties. It is hoped that this will help the current state of ASR development, by adding precision to the discussion of the problem, and by suggesting the direction to possible remedies.

1.3 Thesis Outline

This document begins with an introduction to the motivation and research focus of the study. It continues with a summary of the current speech recognition research and practice, as well as a review of the research into human speech perception and other background literature review necessary for a discussion of this thesis. Next, the document presents the methodology used in the current experiment. The results are then presented and discussed. Finally, conclusions are drawn and recommendations for the application of the findings to future research are suggested.

2. BACKGROUND IN SPEECH SCIENCES

2.1 Automatic Speech Recognition

The current practice of automatic speech recognition is based on computer automated statistical analysis of several acoustic parameters in order to match the sounds to understood words. The process begins with matching the acoustical data of a known text with a direct microphone signal or digital recording device for computer input to train the system. For the system to work properly, the acoustic signal must be in a nearly anechoic setting with little to no background noise. The training process sets up a model of the speech parameters of a specific user, which will be compared to further input for recognition. Statistical methods are used to compare the acoustical parameters of the input speech to the user-specific vocabulary to find the most likely word or phrase being said.

The following description of the speech recognition process is based on Plannerer's textbook on the subject [2]. The process of speech recognition focuses on tracking the important acoustic parameters, the parameters that convey verbal information. Some parameters (e.g. the fairly average fundamental frequency and whether it lies in a man's or a woman's octave range) are irrelevant to decoding the speech signal. Specifically, the short time frequency spectra of the speech signal need to be calculated and tracked as they change from a consonant to vowel to silence. Figure 1 shows a diagram of this process.

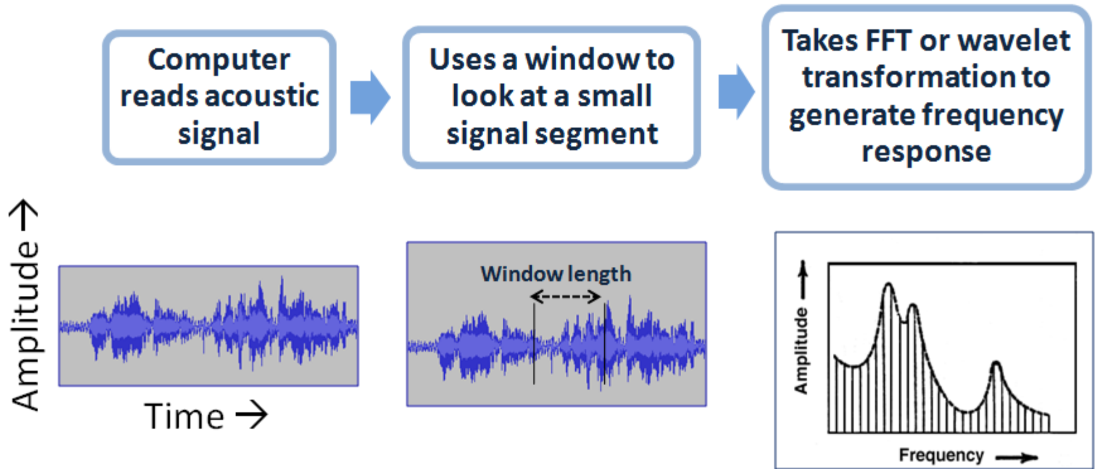


Figure 1: Diagram of Speech Recognition Process

The computer reads an acoustic signal in successive but overlapping time windows to track the speech changes. These windows are typically 10–20 ms long and are varied over several possible locations to maximize their placement relative to the words being decoded. A frequency transformation is performed on each window of data to observe the changing frequency spectrum. A *mel* spectral transformation is applied to produce a perceptually based frequency spectrum. The mel spectral curve applies the logarithmic weighting in Equation 2.1.

$$f_{mel}(f) = 2595 \cdot \left(1 + \frac{f}{700\ Hz}\right) \tag{2.1}$$

This relationship is shown graphically in Figure 2 [2].

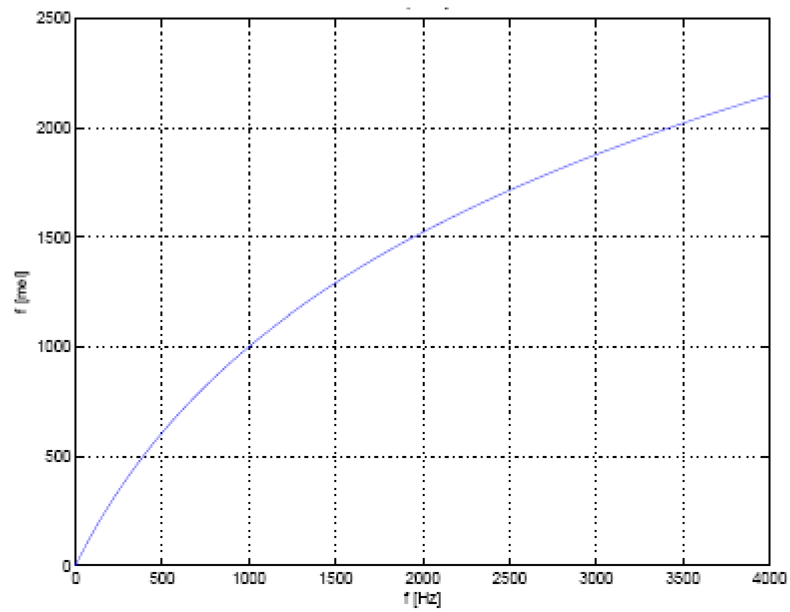


Figure 2: Mel-frequency Scale

The constant fundamental frequency and its harmonics are removed by a process called *cepstral smoothing*, pictured in Figure 3. For speech signals, the fundamental frequency is essentially constant over a 20-ms window period, and the harmonics are consistently related to the fundamental based on the shape of the human vocal tract. The fundamental frequency and harmonics are spaced at integer multiples, (exponentially spaced intervals in the linear frequency domain). In the top of Figure 3, the fundamental and harmonics are easily discerned as the regularly spaced (on the logarithmic mel scale) peaks up the mel frequency spectrum. A cepstral transformation treats the power spectrum (magnitude, not phase) as a time domain signal and applies a Fourier transform, to produce a frequency spectrum of the spectrum, or *cepstrum*. This *quefrency* domain is in essence in the time-domain, though as arrived by this process it does not preserve all phase information. In treating the frequency spectrum as a time domain signal, it turns the regularly spaced peaks of the harmonics into a single peak in the quefrency domain. This isolated peak is the first and largest of the peaks pictured in the middle of Figure 3. A solid line sufficiently above the peak of fundamental harmonic spectral content shows the uppermost cutoff point, under which the spectrum is zeroed out and removed. This is a sharp low-pass filtering of sorts via resetting the Fourier coefficients. Finally, an inverse frequency transform is applied to the cepstrum to produce the smoothed out spectrum pictured in the bottom of Figure 3. The resultant frequency spectrum is a representation of only short term spectral attributes of speech [2]. This spectral magnitude forms the mel-frequency cepstral coefficient (MFCC) at each frequency, in the same way that Fourier coefficients represent linear spectral magnitude.

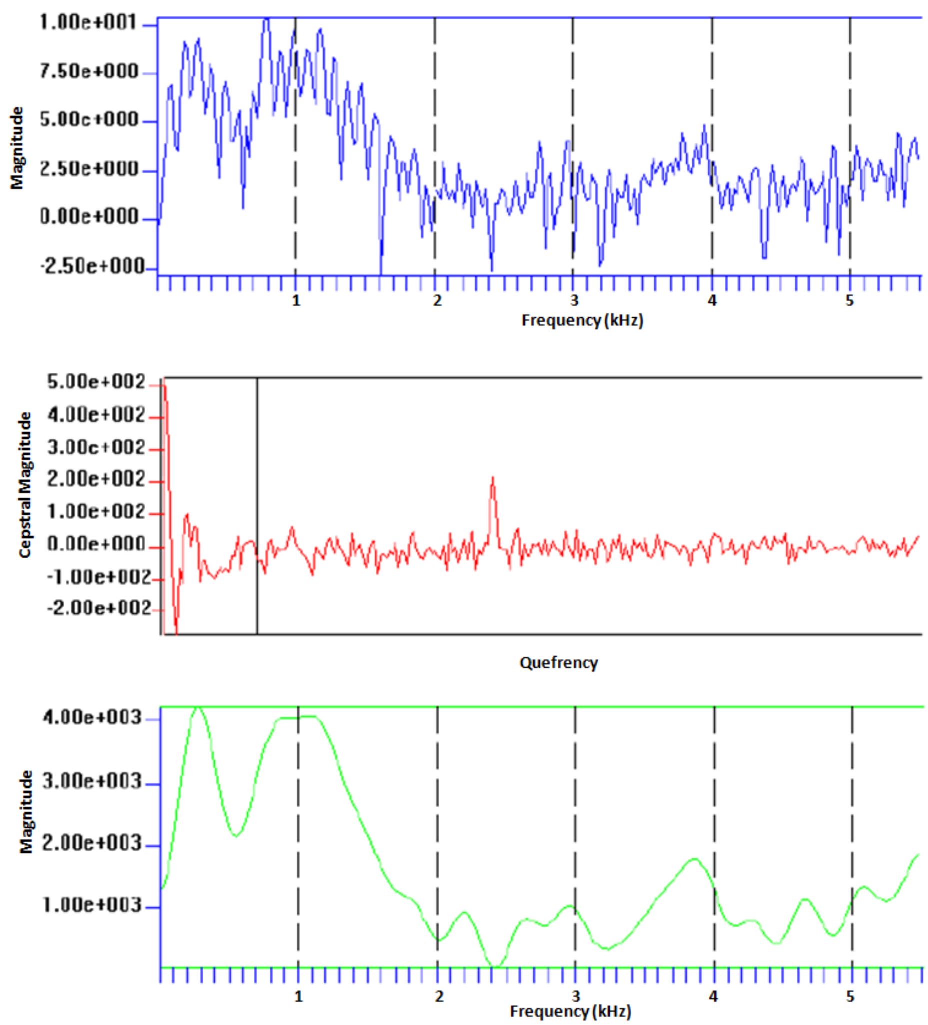


Figure 3: Cepstral Smoothing Process

Top: Power frequency spectrum of vowel "A" before cepstral smoothing. (sampling frequency = 11kHz, number of points = 512). Middle: Cepstrum of vowel results in peak from fundamental and harmonics. Bottom: Fundamental peak removed by high pass and inverse transform to cepstrally smoothed power frequency spectrum.

The acoustical features which are important to the recognition task comprise a feature vector. Specifically, the MFCCs, as derived from the previous process, and their time derivatives are the main parameters of the feature vectors which are calculated for every window to track the speech features. The first order time derivative of evenly spaced windowed coefficients is the difference between time-sequential feature vectors. Each phoneme (the smallest individual unit of speech sound in a language) or word in a vocabulary has a specific feature-vector series in time. The feature vectors track the formants of the vowels, the time-variant presence of white noise (from consonants), and the nature of their combination. The feature vector of the input sound is calculated and compared to the feature vectors of each vocabulary element to achieve maximum similarity or minimum vector space distance between the vectors. The word in the vocabulary which is nearest to the input word in this vector space is the word recognized.

For speech, it is important for a word that is spoken faster or slower to be recognized as the closest to its match in the vocabulary. A matrix technique called *dynamic time warping* adjusts the feature vector comparison to account for the speed of pronunciation [3]. Basically, the time axis of the feature vector matrix allows for horizontal (same feature component, forward movement in time) or vertical (same unit of time, forward to next feature component). Thus a given input sound can travel through the feature changes faster or slower than its vocabulary model as long as every feature change is completed in the same sequence. Figure 4 shows horizontal movement in the comparison between two feature vectors. If [w0, w1, w2, w3, w4, w5] are the feature vector series of a phoneme model spoken at normal speed, and [x0, x1, x2, x3, x4, x5, x6, x7] are a measured feature vector spoken more slowly, this figure shows how allowing for horizontal movement helps to match model with measurement.

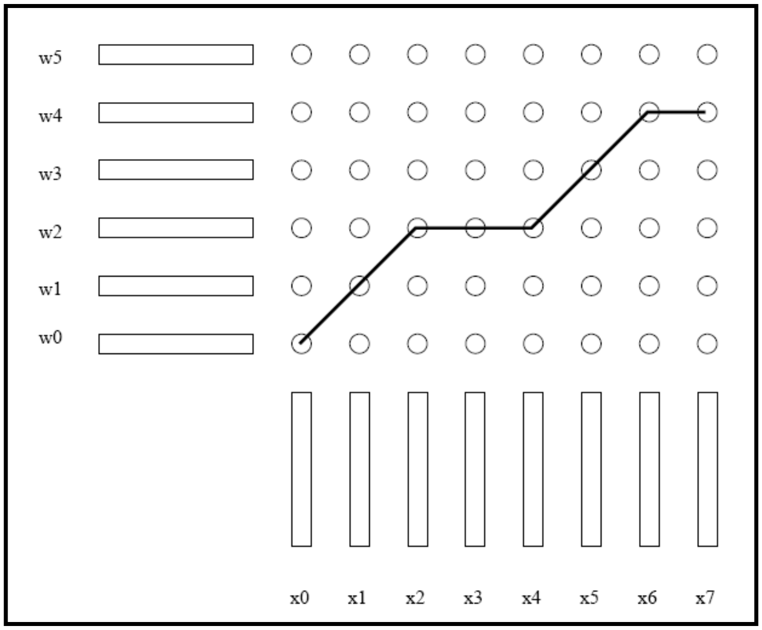


Figure 4: Feature Vector Comparison

Feature vector comparison between phoneme model and measured signal showing horizontal movement. Line shows matching of successive speech elements between a model (y-axis) and measured speech (x-axis). Diagonal portion represents one-to-one movement model and measured speech features, horizontal line represents measured speech which moves forward in time but corresponds to the same model element.

Although the speech recognition system described above is plausible and implemented later in this research, for practical cases of speech recognition this method would necessitate the user producing a model (or ideally several) for every word in the vocabulary. Thus it is necessary in practical situations for the recognizer to come up with its own models of the feature vector for comparison. This process is a *Hidden Markov Model* (HMM) [4]. Markov probability is a process of conditional probability calculation through a matrix of potential conditions. Instead of containing each of the feature vectors of the vocabulary as they change in time, this HMM encodes the probability that each speech element moves to every other speech element in succession.

Markov chains are tables of conditional probabilities which can be used to track any series of conditions to find the probability of an overall result. A typical example is simple weather forecasting. If the *a priori* probabilities of each weather condition moving to any possible weather condition on the successive day are known, they can be listed in a Markov table such as Table 1, where each of three possible weather conditions is represented by a letter—rainy (R), cloudy (C), and sunny (S). Then, the outcome of any particular series of weather events is just the multiplication of the individual conditional probabilities and the overall probability of the initial condition, in Equations 2.2 and 2.3.

$$P(S,C,C,R) = P(C|S) \cdot P(C|C) \cdot P(R|C) \cdot P(S) \tag{2.2}$$

$$P(S,C,C,R) = 0.4 \cdot 0.2 \cdot 0.3 \cdot \mathrm{P(S)} = .024 \cdot \mathrm{P(S)} \tag{2.3}$$

Table 1: Markov Table for Weather Conditions

		Outcome		
		Rainy (R)	Cloudy (C)	Sunny (S)
Condition	Rainy (R)	$P(R \mid R) = 0.3$	$P(C \mid R) = 0.5$	$P(S \mid R) = 0.2$
	Cloudy (C)	$P(R \mid C) = 0.4$	$P(C \mid C) = 0.2$	$P(S \mid C) = 0.4$
	Sunny (S)	$P(R \mid S) = 0.3$	$P(C \mid S) = 0.3$	$P(S \mid S) = 0.4$

Rather than a direct Markov process as shown above, the speech recognizers use a Hidden Markov Model. The Markov probability comparison runs in tandem with the speech feature vector calculation and the probability of each phoneme being represented by the series of feature vectors measured is calculated. The model decides the phoneme with maximum likelihood and returns that as the decision. The process is repeated on a macro level for producing words and sentences.

Figure 5 shows the calculation of the probability of a phoneme. For each successive feature vector, the independent conditional probabilities are multiplied to determine the probability of the series, while each series of feature vectors which arrives at the same phoneme are added. These include series with circular paths, where part of the path is the probability of the feature vector being followed by the same feature vector. Likewise, the probability of a word being said is based on the linguistic conditional probability of a phoneme combination, and the probability of a sentence is based on grammatical conditional probability of a word combination.

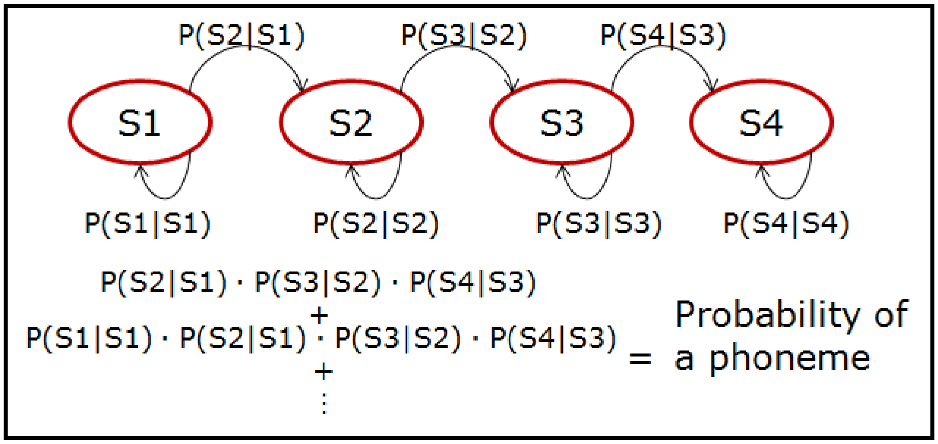


Figure 5: Probability of a Feature Vector Combination

The probability of a phoneme combination is based on the conditional probabilities of the series of feature vectors. S1–S4 are phonemes. Independent conditional probabilities are multiplied for the probability of a feature vector series. Separate feature vector series are added for the probability of a phoneme.

The HMM uses Bayesian statistics to determine the probability that a given word was represented by a set of feature vectors based on the known probability of that word producing the feature vector. The basic formulation of Bayes Theorem is shown in Equation 2.5.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{2.5}$$

In this case, event A represents a phoneme model, and event B represents a specific feature vector measured. The probability that phoneme A is represented by feature vector B, P(A|B), is the quantity which, if maximized, will lead directly to the speech recognition of the feature vector. The three quantities on which this probability is based are the conditional probability of the feature vector given the phoneme, P(B|A) (as determined by the recognizer's vocabulary model and user training), and the overall probabilities of the phoneme P(A) and the feature vector P(B) being produced in general. In this case, the quantity P(B) is unknown but unimportant to the maximization task, since the feature vector under question is a constant for all speech recognition tasks [5].

Thus, by maximizing the conditional probability quotient, the HMM determines the most likely phoneme that the feature vector represents. Likewise, HMM models and Bayesian statistics allow for word and grammar models at more macro levels. Taken as a whole, this process allows the HMM to form a speech model for a large vocabulary of words based on a small amount of training material, and is the most common implementation of speech recognition today.

2.2 Human Speech Perception

In many ways, the study of human speech perception encompasses and predates the study of automatic speech recognition. The processing in speech recognition technology is based on human speech perception, and significant parallels exist. In other ways, however, the field of human speech perception has made new advances since the creation of automatic speech recognition. The ASR world has largely branched off to optimize implementation of the chosen statistical methods outlined above. At all times, computational limitations of the processing have limited the ability of ASR to incorporate the full extent of the knowledge base of the human processing of speech.

Speech intelligibility is an empirical measure based on correct identification in subjective word recognition tasks. In practice, it is most commonly reported by the Articulation Index (AI), a metric first developed at Bell Laboratories to test the transmission of phone lines [6]. In order to create a predictive and objective measure of this subjective quantity, there has been a large research effort to define and study room acoustics metrics which correlate well with speech intelligibility (most of which, by the way, are unmentioned by the ASR community). In her master's thesis, Dorea Ruggles, provides a concise review of the current speech metrics, their definitions and relationships to intelligibility, and the following summary of intelligibility metrics comes directly from her research (Ruggles, 2007). The speech intelligibility metrics can be divided into three basic categories, based on: 1) acoustical energy ratios, 2) reverberation times and signal-to-noise ratios, and 3) modulation transfer functions.

Acoustical energy ratios, the most popular of which are clarity (C50 or C80) [7] and definition (D) [8], are a common way to estimate intelligibility. Clarity is the decibel expression of the fractional comparison of the impulse response energy in the first 50 or 80 ms to the part of the impulse response after that time. The formula for either C50 or C80 is shown in Equation 2.6, where x is either 50 or 80 ms and h(t) is the impulse response of the transmission path being analyzed.

$$C_x = 10 \cdot log_{10} \left| \frac{\sum_{t=0}^{x} h(t)^2}{\sum_{t=x}^{\infty} h(t)^2} \right| \tag{2.6}$$

Definition is the fractional comparison of the first 50 ms of energy to the total impulse response energy, shown in Equation 2.7.

$$D = \frac{\sum_{t=0}^{50ms} h(t)^2}{\sum_{t=0}^{\infty} h(t)^2} \cdot 100\% \tag{2.7}$$

Since early energy is most important to speech intelligibility, the C50 and D metrics are generally used for reporting on speech, whereas the C80 metric is used more for musical settings. Although they are strongly correlated with each other and slightly less so with reverberation time, the basis of clarity and definition on a perceptually important distinction between early energy and late (reverberant) energy has been shown to be useful in predicting intelligibility. These metrics are fairly simple however, and cannot take into account disturbances which are more complex and not represented by the overall energy balance (including steady-state and impulsive noise).

Reverberation time alone has been a good measure of speech intelligibility, similar to energy balance measures. Recent research also shows that concurrent analysis of the frequency specific reverberation time and signal-to-noise ratio (speech volume to background noise) can be combined into an equation which mathematically predicts the articulation index to fair precision [9]. Although the specific mathematical combination of these two separate metrics is still debated and not widely used, it seems that these two metrics separately are two of the most accessible and most widely used metrics to describe acoustic settings outside of the acoustics community.

A final branch of speech intelligibility metrics are based on modulation transfer functions. These metrics evaluate the change in a stimulus signal that the transmission path has affected vis-à-vis a filter on the original speech envelope. Modulation transfer functions have the potential to account for a combination of reverberation, noise, and other distortions (a notable advantage—those distortions which do not qualify as either of the previous categories) into a single measurable quantity. The most common of the modulation transfer function metrics are speech transmission index (STI) and rapid speech transmission index (RaSTI) [10]. Their reported variability of up to 5-6% standard deviation has made interpretation of their results somewhat questionable in different settings [11]. Still, since the early 1990's, modulation transfer functions have been incorporated into computer modeling of architectural acoustics and are widely used in both the research and acoustical design communities as reliable estimates of speech intelligibility. Ruggles' research has added to the discussion by incorporating binaural measurement into the calculation to account for this aspect of human hearing, a step which could greatly increase the relevance of these metrics [11].

Binaural recording and measurement is a topic which has begun to be addressed by ASR researchers (see Section 3.2), but binaural human hearing has been studied in some depth for many years. Many of the findings from the research on binaural hearing have implications on ASR systems. The "Cocktail Party Effect" allows binaural listening cues on directionality to segregate signals which are indistinguishable from monaural listening or omni-directional recording [12]. The mechanisms for binaural processing include inter-aural time differences and inter-aural level differences. The equalization cancellation theory (whereby the binaural system cross-correlates the signals to determine the coherent signal intended for detection) helps explain the brain's remarkable ability to distinguish a speech signal from noise [13]. Reverberation is directly at odds with this discrimination method, since reverberant energy is more correlated with the direct sound and more correlated from ear to ear than randomly fluctuating background noise. Lavandier and Culling [14] propose this theory as an explanation for why strong late reverberant energy and echoes are the most damaging components of an impulse response transmission path.

Another significant unresolved problem for ASR is the recognition of a target speech versus competing background speech with similar or unpredictable qualities. Culling and Binns [15] have also investigated this problem in the human realm—speech reception thresholds for intelligibility against competing speech are inexplicably high. Their theory for the basis of the brain's ability here is its ability to detect target-to-interferer ratios in specific frequency ranges which are produced by head-related transfer functions, and which allow the brain to suppress the competing speech. In headphone conducted experiments which eliminate head-related transfer functions, when elements of these frequency dependent head-related amplitude attenuations were tested, they had a dramatic effect on the detection and intelligibility of a target speech source. Previous research had also shown a specific dependence on the frequency shaping of the fundamental versus the mid-high frequencies (2–3 kHz) commonly important to speech intelligibility.

There are significant developments from the recent speech sciences research that have promising implications on the improvement of ASR. There is an understandable lag in the adaptation of the current research into ASR technology, as is noted in the upcoming section. Nevertheless, the potential for improvement of methods stands such that it is important for the ASR community to stay abreast of the current speech sciences research.

3. REVIEW OF CURRENT ASR RESEARCH

Given the limitations of automatic speech recognition relative to human speech perception in noisy and reverberant settings, the current ASR work focus is on improvements in these two settings. Although several new strategies have come out of this research, the research is widely focused on implementation development and there is comparatively less time spent on analyzing recognition performance of various implementations to better understand the problem. Also, several studies within the field could be more precise in their discussion of room-acoustics science and do not test a very large breadth of room-acoustics settings.

3.1 Techniques for Improving Speech Recognition in Noise

In the speech recognition research, missing information refers to all cases in which the speech signal is masked by some corrupting signal, and thus encompasses both noise and reverberation masking. The area of speech recognition in noise has been widely explored and significant advancements have been made.

The main parameter for determining speech recognition accuracy in a noisy environment is the signal-to-noise ratio. Additional factors contributing to accuracy include the frequency content and time domain transient nature of the noise signal. For signals with a consistent and predictable noise signal, simple pre-processing techniques such as filtering can increase the signal-to-noise ratio to achieve a near perfect recognition rate [1]. For signals with variable and unpredictable noise corruption, advanced techniques are necessary. One such technique for separating the noise signal of a competing talker is described by Yen et al. [16]. Their paper describes a process for front-end adaptive decorrelation of a dual channel system, one microphone for each speech signal (but with simulated crosstalk from close proximity). The experiment used 78 sentence pairs from an ASR testing database (TIMIT) and a vocabulary size of 835, recorded anechoically. The crosstalk between proximal speakers was simulated at 1 meter distance to each other and 10 cm from their sound sources. For five levels of relative source energy levels (RSEL) between the source and interferer, the drop in recognition accuracy from that of the original signal is from 91% to 20-60%, depending on the RSEL. Their adaptive decorrelation algorithm combined with time-domain and frequency-domain signal detection bring recognition levels back to within 10% of the original recognition accuracy for all RSEL above –20 dB.

Also tackling the problem of a competing talker as a noise source is a method of determining usable speech content proposed by Iyer et al. [17]. The experiment tested a small vocabulary of 10 utterances, but with 48 different male and female speakers. The test signals were juxtaposed to combine signals from separate speakers. The identification algorithm found the feature vector space distance from the training material to identify and recognize only the usable speech. The identification algorithm also used variations of logic trees to decide which speech elements to focus on in the comparison. Correct identification rates were 65–70% depending on parameters, and false alarm rates (recognizing the wrong speech text) were approximately 20%. Another study uses the determination of usable speech as a technique for improving speech in the presence of variable noise environments [18]. The study employs modulation filtering to determine the parts of the signal which vary with speech properties and those which are contaminated with noise or reverberation.

Physiologically based techniques for speech-in-noise recognition represent a promising research initiative in the field. The research premise for Holmberg et al. was based on noted similarities between the first stages of the auditory process and the speech recognition algorithms [19]. The study endeavored to include a correlate of the short term adaptation of a feature model based on physiological synapse properties. They tested with both additive Gaussian white noise, and the AURORA 3 test, which tests digit strings recorded in a car under various driving conditions. The results were averaged over four different languages. For their synaptic adaptation models, they experimented with an adaptation time constant, which they showed experimentally to be optimized between 200 and 300 ms. They note that this result aligns with a physiological argument for approximate human forward masking recovery time (> 200 ms). Their results show robustness towards steady-state noise sources which can affect feature models, similar to the brain's ability to filter out a constant background noise.

Another physiologically based technique is the concept of auditory scene analysis, through which a human processes and understands the signals presented to him. A study by Brown et al. combined a neural oscillator technique for auditory scene analysis to parse the signal into speech and noise, with a missing data recognition algorithm for improvement of the speech signal [20]. Their missing data recognition algorithm involves a parallel calculation of the reliability of a speech segment to weigh high quality speech samples higher than low quality speech samples in the recognition process. In testing a vocabulary of 240 male speech digit utterances, recognition accuracy was 10–30% more accurate than conventional ASR methods (HMM processing and spectral subtraction) at low signal-to-noise ratios.

Binaural processing is another way that humans are better able to filter out noise than a monaural system. In their paper, Palomäki et al. showed that by applying binaural geometric acoustics, including low order wall reflections, and source identification algorithms they could improve recognition rates by targeting the directional speech source and attenuating the omni-directional noise source [21]. Their algorithm increased accuracy 40% for a -5 to 30 dB SNR range, though still only achieved 70% accuracy when a small reverberation is included (0.5 s reverberation time). This technique of course requires a precise knowledge of the room characteristics and receiver position ahead of time and might not be as practical in a real world technology. It is, at least, robust to changes in speaker location as it uses a reflection discrimination algorithm to map the source in any position.

3.2 Techniques for Improving Speech Recognition in Reverberation

In general, the focus on improving speech recognition in reverberation has been a comparably endeavor. The research effort started gaining traction in the late 1990's. One of the first comprehensive studies was performed by Kingsbury in his doctoral research with several signal processing strategies developed and implemented [22]. He investigated the robustness of ASR speech features to reverberation and noise when preprocessed with FIR filtering via approximate critical bands. He also experimented with the RelAtive SpecTrAl (RASTA) method, which tracks slow changes in the spectral structure of the speech signal relative to its average, and found robustness to linear spectral distortions created by room effects. Additional techniques implemented were envelope windowing and automatic gain control. Each of the optimized techniques has a demonstrated improvement, though some of the techniques such as automatic gain control have not had much prominence in the further research.

One prominent research focus is deconvolution with the room impulse response. With a basic knowledge of the source/receiver path, the reverberant energy can be removed from the received signal (theoretically) to its original clean form. The challenge with this method is to develop a system which is robust to slight changes in the source/receiver path (e.g. the talker moving his head position) and to have the system responsive without prior knowledge of the room (which is the case in many settings). It would be ideal for the system to recognize the source/receiver path out of the received output and to apply the appropriate algorithm to remove the reverberant energy without intensive input from an engineer. This dereverberation by so-called "blind" deconvolution appears to be beyond the reach of the current state of the art. As suggested by Hatziantoniou et al. [23], the overall reverberant behavior of a room could be measured in situ by a speaker/microphone impulse response measurement, such that at that position exactly an automatic speech recognizer would have near perfect performance. The extrapolation to another position in the room would be of variable success. However, even this idealized deconvolution theory is unpractical. A room's acoustic response is well modeled by a finite length impulse response, i.e. an FIR filter, and the precise inverse filter of an FIR filter is an infinite impulse response (IIR) filter. Thus to perform the deconvolution ideally, it must be completed offline. Hatziantoniou et al. developed a technique for producing a complex smoothed transfer function, which can be translated into a finite approximation of the inverse room impulse response. This pre-processing technique has shown small but consistent improvements in accuracy (< 5%) in both a classroom and concert hall setting.

In the same vein, Park et al. developed a two microphone dereverberation technique which incorporates previous knowledge of signal and interferer positions to suppress the reverberant energy and noise [24]. The algorithm uses temporal suppression of interference signals (similar to the brain's processing using the precedence effect) for a speech recognition improvement of 10–20%. In combination with directional suppression (based on signal/interferer known positions and an interaural level difference between microphone signals), the suppression processing algorithms can achieve a 60% accuracy improvement over the original signal. A similar research effort was started prior to Park et al., with a two microphone segregation technique that also uses directionality information a priori [25]. Using temporal cues to imitate the precedence effect and harmonicity cues of the voice signal to suppress non-harmonic noise, up to 70% accuracy improvement was found at 0 dB signal-to-interference ratio.

Several other monaural techniques have been proposed which do not require a previous knowledge of the source/receiver path and therefore are more practical methods for ASR improvement of the current technology. The method proposed by Gillespie et al. is to partially identify the impulse response using the correlation of multiple microphones and then to shorten the impulse response by cutting off the reverberant tail [26]. It was found experimentally through human testing that the information at the extremes of the cross-correlation were unrelated to the speech sample and unimportant to the recognition task. This information was then removed, thus shortening the impulse response. The ASR accuracy improvement was around 10%.

Another blind reverberation improvement method is long-term spectral subtraction, developed especially by Gelbart et al. [27]. In calculating the long term frequency spectrum of a talker in a room, the average shape of the reverberant energy response can be acquired with a set range of variability of the speaker spectrum. This study showed that a log subtraction (mathematically interchangeable with division) of the reverberant speech spectrum produced an increase in accuracy in reverberant settings of 10 to 40%. Gelbart et al. later expanded on their original research to a far-field conference room application [28]. Here a combination of noise reduction and the log-spectral subtraction method showed an improvement in word error rates of about 20%.

Takiguchi et al. incorporated first order frame-by-frame prediction model into the model to account for the building up of a reverberant field [29]. The model takes into account the speech being received and uses a running tabulation of the average room effect to predict the specific room effect in the coming frame. The model defines the predicted room effect in two parts: a spectral shift from discrete reflections and an additive noise signal from reverberant energy. These predicted effects are accounted for in the HMM model. The system was tested in a reverberant setting while varying the speaker/receiver distance, thus changing the direct to reverberant ratio. For the settings tested, the recognition system shows no degradation of recognition accuracy from 90% original. A related study worked to develop a clean model of the preceding frame in order to provide a more accurate prediction of the reverberant effect in the current frame [30]. The study confirmed an experimental improvement in accuracy based on isolated word recognition.

Another research initiative is the incorporation of binaural hearing models into ASR, since it is well known that humans hear better binaurally than monaurally. Several of the two microphone techniques discussed above have been applied with a binaural model including head-related impulse responses (HRIRs) [31]. The algorithm uses HRIRs to create a binary classification of sound signals into foreground and background. For all tests, the target is fixed at 0° azimuth, while the interferer is moved in azimuth and type of signal. The target is therefore known in advance, but recognition improvements are exhibited for a range of interference signals. A real-world variable interference would be more problematic since the adaptation suppression process takes a few seconds to recalibrate itself. Some processing time could be reclaimed using an engineering solution (i.e. a dummy head) rather than an additional processing step for applying the HRIRs, though this would also require a revamping of the current state of the art technology.

Although this sampling of the reverberant ASR research is not fully complete, it represents the range of strategies currently being implemented in the ASR research community. While there may be developments which precede or derive from the studies or represent variations on the strategies mentioned, they will largely fall into one of the categories of research above, that is, binaural and two microphone measurements, blind or deconvolutional dereverberation, or model training for reverberation adaptation.

3.3 Comparative Analysis of Recognition Techniques

Upon review of the literature, it seems that one of the limitations of the current field is an abundance of technique development with little precise comparative analysis of the methods' efficacy at solving the technology's problems. By contrast the focus is largely on developing a new method and reporting the method's accuracy in a limited number of settings. There are not enforced standards for speech recognition tests, and there are so many variations of source material, interfering signals, and recognition parameters. As a result, separately reported results are often incomparable. What would be useful in response to these diverse methods and findings, is a concerted effort to implement and test various techniques in a standardized way, and report and analyze this comparison in a meaningful way. Unfortunately, in the past decade, only a handful of researchers published papers with a comparative, analytical focus.

Milner implemented and compared performance of several front-end processing strategies for robust speech recognition [32]. He compares MFCC to perceptual linear prediction (PLP); RASTA to cepstral mean normalization (CMN); and temporal derivatives to cepstral-time matrices (CTM). He produces 10 combinations of these variables and tests their recognition accuracy on a 2560 sentences from the BT subscriber telephony database. Overall, the best performance (46%) was from MFCCs, filtered by RASTA, with CTM as the speech feature for recognition.

Stern, et al. analyzed three signal processing recognition improvement techniques as they compare to typical HMM recognition [33]. The techniques compared fall into three categories: acoustical pre-processing (linear adaptation model to reverberation), microphone array processing, and the incorporation of physiologically motivated models for the later stages of the hearing process. The results show substantial gains from the signal processing technique and microphone array technique individually, though little additional gain from their combination. The physiologically motivated model actually decreased recognition performance from the baseline MFCC and HMM algorithm.

Deng et al. [34] performed an analysis and comparison of two speech feature extraction/compensation algorithms: feature-space minimum phone error (fMPE) and stereo-based piecewise linear compensation for environments (SPLICE). The first technique is primarily used for conversational speech, while the second technique was developed for robustness to noise. Each technique performed best in its designed setting, though the fMPE outperformed the SPLICE algorithm in general.

One study compared three overall platforms on Thai speech recognition [35]. Unlike most studies which take HMM processing as their basis, this research compares the

HMM with the Modified Back Propagation Neural Network, and the Fuzzy-Neural Network. The field of Thai language speech recognition is in a more preliminary stage than English ASR, and presents new challenges due to its tonal nature. The study shows that HMM processing is likely still the best choice of current implementation options for the development of tonal Thai ASR.

Another study conducted a comparison of the performance of several signal processing methods on Chinese language speech recognition [36]. The techniques compared were dynamic spectral warping (DSW), dynamic time warping (DTW), HMM, and learning vector quantization (LVQ2). DSW is a method to allow for a warping of the frequency domain due to spectral speaking variations, similar to the way time warping allows for variations in the speed of speaking. The results show DSW in combination with HMM to be robust to speaking variations at around 91% correct recognition.

Finally, Flynn and Jones conducted a comparative study of auditory-based front-ends for robust speech recognition using the Aurora 2 database [37]. The two front-ends discussed are Perceptual Linear Prediction (PLP), a linear prediction model employing critical bands, and an auditory model (PEMO), a computational model of the auditory periphery. Each of these is compared with HMM processing without front-end preprocessing.

Although more work will be needed to bring the disparate branches of the ASR research together towards a cohesive and standardized field, the researchers cited here have recognized this need and taken a step towards solving the problem. Unfortunately, as noted, each piece of research was only able to implement a few methods for comparison. Admittedly, one obstacle for comparative study is the great challenge in reproducing the techniques of another piece of research. Towards this end, standardization of testing methods would be a good intermediary step.

3.4 Treatment of Room Acoustics within the ASR Literature

In addition to the lack of comparative studies in the field, one of the main problems with addressing the area of ASR in reverberant settings is the lack of precise room acoustics science in the discussion. As Palomäki et al. noticed in his introduction (p. 1), it is "apparent that few computational approaches to ASR have been evaluated in reverberant conditions, presumably because of the difficulty of the task" [21]. Another potential cause is the chasm between room acoustical science, the speech sciences, and the computer science of automatic speech recognition development.

One study which tries to directly take on this problem is Pan *et al.* white paper on how room-acoustics settings affect mel-frequency cepstrum coefficients, critically important to most speech recognition systems [38]. This is possibly the only example of a focused study in the field on the character of the degradation of speech recognition by reverberation. The study found that stationary background noise produces low frequency masking of MFCCs, while reverberation effects has its most deleterious effects on high frequency MFCCs.

Although many of the studies cited previously have documented an improvement in reverberant speech recognition based on their implementation strategy, many of the studies have not included a precise or thorough exploration of the effect on their developed techniques from basic reverberation parameters like reverberation time or room material properties. The following is a review of the experimental setups of the tests, the room acoustics parameters in their methodology, and an attempt to compare their results.

Several of the papers do not give enough information to draw conclusions on the type of reverberation added. Takiguchi et al. only cite the source of the impulse response used and lists its length, not giving any other relevant information on its properties [29]. The room acoustics-focused study by Pan mentioned previously in this section, while precise in its analysis of the effects of reverberation on MFCCs, only uses one reverberation setting in its methodology and only says the testing facility has "moderate reverberation" [38]. Shamsoddini lists the dimensions of the test space and shows the impulse response, but does not analyze it or describe the materials for any further inferences on the reverberation tested [25]. Without reporting the reverberation time, or a combination of room materials and dimensions, it is hard to be sure of the testing conditions for these studies. This makes it difficult to repeat the methodologies to either confirm or build on the findings of these studies.

Many papers do list either the reverberation times or room parameters of their testing setups, however many of the tests use only a limited range of reverberant settings and do not make a controlled variation of the reverberant setting a primary focus of the methodology. Gelbart et al. test the HTK toolbox HMM recognizer with two reverberant impulse responses, one a convolution with an artificial RT of 0.5 s, and the other a measurement in a room for "natural meetings" which did not have any parameters or measurements reported other than microphone distances [27]. (Unless otherwise noted, all values listed for reverberation time in the literature were reported with unknown frequency content, and must be assumed to be broadband frequency RTs.) Their baseline tests for 0.5 second artificial reverberation show a Word Error Rate (WER) of 19.2%, or an accuracy of 80.8% in recognition of digit sets. Gillespie's research on dereverberation shows the accuracy results for 6 reverberation times as they have removed some of the uncorrelated non-speech energy [26]. With a baseline of unprocessed 0.31 s RT, they measure 40 and 42% (for Microsoft and IBM recognizers respectively) accuracy for untrained large vocabulary recognition (as depreciated from 58 and 66% accuracy for the anechoic signal). The Park et al. binaural dereverberation research analyzes the effect of 0.3 and 0.5 second reverberation times, and signal to interferer ratios (SIR) of 0 and 10 dB [24]. Although the results of the 0 dB SIR tests show a curiously unexplained 107% WER, the results at 10 dB SIR are 57.1 and 71.6 WER for 0.3 and 0.5 s RT respectively. The tests were performed on sentence recognition of sentences which came out of training material, and the SPHINX-III recognition system was used. Hatziantoniou tests the HTK toolbox recognizer with two real room settings measured directly with impulse response techniques in a classroom (RT = 1.1, V = 200 m^3) and concert hall (RT = 2.1, V = 9633 m^3), with recognition rates of 40-50% in the classroom and 10-12% in the concert hall [23]. Roman's tests use 0.1, 0.2, 0.3, 0.4, and 0.5 second reverberation times to analyze the performance of their binaural segregation algorithm on suppressing noise in reverberant settings [31]. However their methodology does not explore the effect of the reverberation times in the main ASR experiment, and mainly focuses on a controlled variation of signal-to-noise ratio metrics. These several papers have taken steps to address room acoustics in their methodology, but the differences in and sometimes incomplete description of the reverberant conditions of their testing methodologies still make the results hard to compare.

Palomäki et al. have included a thorough evaluation of their methodology over a variety of room acoustics environments [21]. The precision of their discussion and breadth of testing of room acoustics parameters might be a good model to look to for testing of reverberant ASR. Their research uses an image-source method for modeling the impulse response of a small office-like room, with dimensions 6 by 4 by 3 m. The binaural listener was placed in the artificial model at 3 by 2 by 2 m as measured from the same origin as the dimensions are reported. They tested the algorithm (which groups sounds by a common azimuth for noise and reverberation masking) with 8 wall reflection factors: 0, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, and 0.95. The results with no processing show a decline in recognition accuracy from increasing the wall reflection factor (decreasing absorption). Though he does not state explicitly whether this wall reflection factor is a pressure or an energy ratio of reflected sound, his statement that the reverberation time of 0.9–0.95 corresponds to a reverberation time of around 0.5 s suggests mathematically that he refers to a pressure reflection factor, which is the standard for reporting a wall reflection factor. The wall reflection coefficient reported in his study corresponds to the absorption coefficient in this study by Equation 4.1,

$$\alpha = 1 - |R|^2 \tag{4.1}$$

where α = energy-proportional absorption coefficient and R = pressure-proportional complex reflection coefficient.

The results of the Palomäki et al. study are pictured below, and the results of a simple reverberation time of their stated room-acoustics parameters are listed in Table 2.

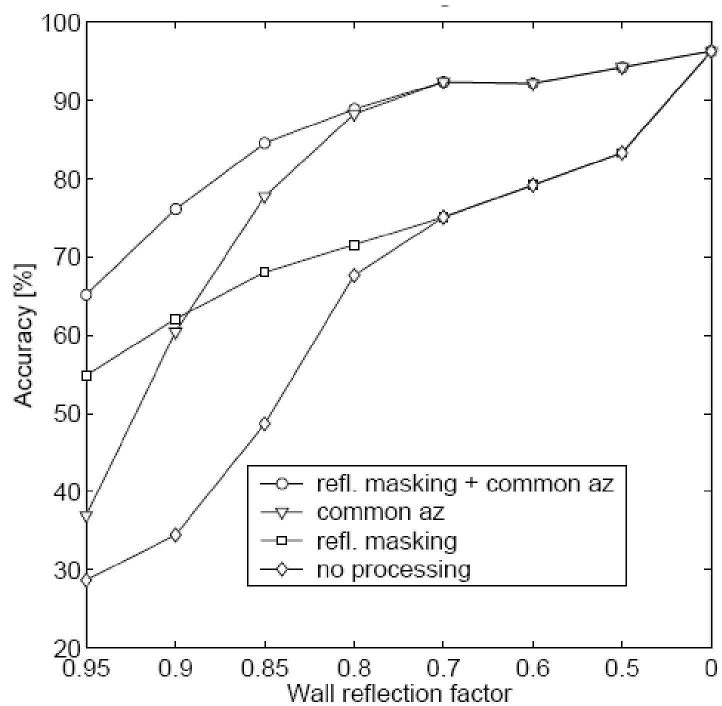


Figure 6: Palomäki et al. Results for Accuracy versus Wall Reflection

Results of Palomäki et al. study taken from their 2004 JASA paper. Graphed in connected diamonds is the original processed reverberant ASR performance.

Table 2: Reverberation Time Calculations for Palomäki et al. Data

Wall Reflection Factor (R)	0.95	0.90	0.85	0.8	0.70	0.60	0.50	0.00
Absorption Coefficient (α)	0.10	0.19	0.28	0.36	0.51	0.64	0.75	1.00
Reverberation Time (Sabine)	1.11	0.57	0.39	0.30	0.21	0.17	0.14	0.11
Reverberation Time (Eyring)	2.44	1.19	0.77	0.56	0.35	0.24	0.18	∞

As shown above, there is an issue of breadth and precision in the application of room acoustics simulations to ASR testing procedures. Although some researchers perform a full room acoustics evaluation of their experimental setups, there is no standardization for reporting reverberation times, room material properties, or other room acoustics parameters. There was not found in any paper any mention of clarity, definition, or speech transmission index, which have been found to be helpful parameters in the human speech sciences community. (Although definition and clarity may have limited meaning because they are largely derivative from reverberation time, they could still be useful to the discussion of automatic speech recognition as they have been useful in speech sciences.) The purposeful inclusion of more room acoustics parameters in the discussion of ASR development would not only help the discussion to be more informed of the principles behind the problem at hand, it would help researchers compare their results to one another and allow them to reproduce the experimental conditions.

3.5 Contribution of Current Research Effort

This research effort will add to the current ASR research literature by performing a limited number of comparisons of common speech recognition implementations in a variety of room acoustics settings. Although the research will not endeavor to find new implementation strategies, the results should increase understanding of the problems in reverberant ASR. Although the room acoustics science will be treated with the precision of the room acoustics field, the speech recognition science will be largely taken for granted. That is, this research will mainly treat ASR systems as a black box in themselves, in which the system identification task at hand is to quantify more fully their performance with respect to different amounts of reverberation. Although the implementations chosen are not at the cutting edge of the field, hopefully the findings will point to a deeper understanding of the problem at hand, just as the model of evaluation of reverberant speech recognition can help inform the process of evaluation for those at the forefront of the field.

4. EXPERIMENTATION

4.1 Methodology

The methodology of this experiment entails imposing simulated room-acoustic situations on a set of input speech samples. The speech samples are run through an ASR platform and the average accuracy is determined over each room acoustic parameter.

4.1.1 Automatic Speech Recognition Platform

Several ASR platforms were explored in the preliminary stages of the thesis. One platform, which was explored but not fully developed in this experiment, is HTK Toolbox, a C++ based research tool developed by researchers in Cambridge, UK. It gives the user full control over the ASR process including testing and training materials, vocabulary, and HMM parameters, but it therefore requires a large amount of study for successful implementation [39]. Dragon Naturally Speaking® is a commercial speech recognizer with built-in HMM recognition [40]. The Dragon platform provided ease of setup, and confidence that it was operating in a typical way (for commercial speech recognizers), but was limited in its research capabilities. Dragon has suboptimal individual word recognition, since it is mainly intended for continuous speech use. It is also limited in its flexibility for atypical setups such as limited vocabularies and user-defined training materials. The other platform for the thesis methodology is a Matlab toolbox developed by Luigi Rosa [41]. It uses MFCC feature vector comparison with dynamic time warping to perform single word recognition. While not precisely an HMM recognition process, it is a simplification of the process which will provide a more direct evaluation of how speech features are degraded in reverberation. The Matlab toolbox also provides the flexibility to adapt the open-source program into a specific implementation.

The two chosen recognizer platforms were the Matlab toolbox and Dragon Naturally Speaking. This is a first step into exploring the exact nature of the effect of reverberation on speech recognition platforms; the basic experimental method outlined in Section 3.1 could easily be applied to other platforms in the future.

4.1.2 Pilot Test

To test the functionality of the recognizers, each was evaluated in its typical setup. A preliminary test on the Matlab recognizer was performed, using the numbers 1 through 10. The recognizer was trained with one set of the numbers spoken (i.e. this was its vocabulary), and its recognition was tested with 4 different sets of the numbers spoken. In other words, this tested the recognizer in its typical use, with different speech files for training and testing, and with a vocabulary of very different speech files. The Matlab recognizer had 100% accuracy at this test, and thus was shown to be effective at a very typical task. The general setup for the thesis as described later uses the recognizer in an atypical manner, so having the basis of a typical performance is an important starting point.

Likewise the Dragon recognizer was tested with continuous speech, to ensure proper installation and functionality. The recognition performed at a reasonable level, however a precise accuracy statistic is difficult to report for a continuous recognition task.

4.1.3 Source Material

The source material was a list of 300 words from the modified rhyme test (ANSI Standard S3.2 1989, see Appendix 0). The words are in 50 sets of 6 similar sounding words, which are typically presented to a human subject in a specific setting to determine the intelligibility of the speaker/listener communication path. The subjects are given the set of words as the choices for identification of each word said, and the intelligibility is determined by how far above random guessing (1/6 accuracy) the subject accurately identifies the word spoken. The speaker/listener communication path can contain an air path, electronic path, and/or visual cues. It is usually evaluated in comparison to a setting where the speaker is sitting in front of the listener so all paths are optimal. This is a 100% recognition benchmark to evaluate by, in case there is confusion due to pronunciation or identification. The sets of 6 are similar one syllable words, starting and ending with a consonant, with a vowel in the middle. The sets have either the same vowel and ending consonant (differing in the starting consonant), or the same starting consonant and vowel (differing in the ending consonant). For example, the following represents a first category set: went, sent, bent, dent, rent, and tent. This organization limits the user choices and allows the researcher to focus on a micro level at exactly which speech elements are being misidentified. For thoroughness in this study, 6 new sets of 6 words were added which had the same starting and ending consonants but differed in the vowel, although this is admittedly a smaller sample size than the 25 sets of each other case.

Although the standard specifically says it does not necessarily apply to automatic speech recognition analysis, it is a good starting point. By training the recognizer with each set and then testing with each word of the set, the test can analyze how degradation of the speech features results in recognition of a similar sounding word. This list works well towards a targeted analysis of small phonetic changes within each set, but with broader scale statistical implications from averaging a significant number of recognition tasks together.

All samples were recorded in the Jaffe Laboratory in the basement of the Greene Building using an MXL 990 Large Diaphragm Condenser microphone, a Creative USB Soundblaster (Model No. SB0490), an M-Audio AudioBuddy Dual Mic Preamp/Direct Box, and a Stagg Popscreen (Model No. PMCOH). See Appendix A for pictures of the recording setup.

4.1.4 Mirror Room Impulse Response Modeling

The room acoustic settings for this experiment are modeled using a geometrical acoustics platform implemented by Braasch [42]. This Matlab toolbox, “Mirror Room,” uses source mirroring about the walls of rectangular rooms to perform an image source model of the early reflections. The model specifies room size, source and receiver positions, and wall and floor material absorption coefficients (in octave bands). In its current form the mirror room program uses an omni-directional source and a binaural receiver oriented towards the source. Each source is mirrored about the boundaries of the ideal room for each additional order reflection, for the first 1000 mirrored sources. The Head-related Transfer Functions (HRTFs) then filter the sources for arrival at a specific degree relative to the binaural receiver. The HRTFs used to create the impulse response have a resolution of 1° azimuth and 10° elevation.

The procedure for creating source material was to convolve impulse responses having a series of varied room acoustic parameters with anechoic recordings of the modified rhyme test material, in order to model the recognition of speech in a wide range of rooms. The impulse responses used had a room size of 4 by 5 by 3 meters, a source position of (2, 1, 1) m, and a receiver position of (1, 1, 1) m, as shown in Figure 7.

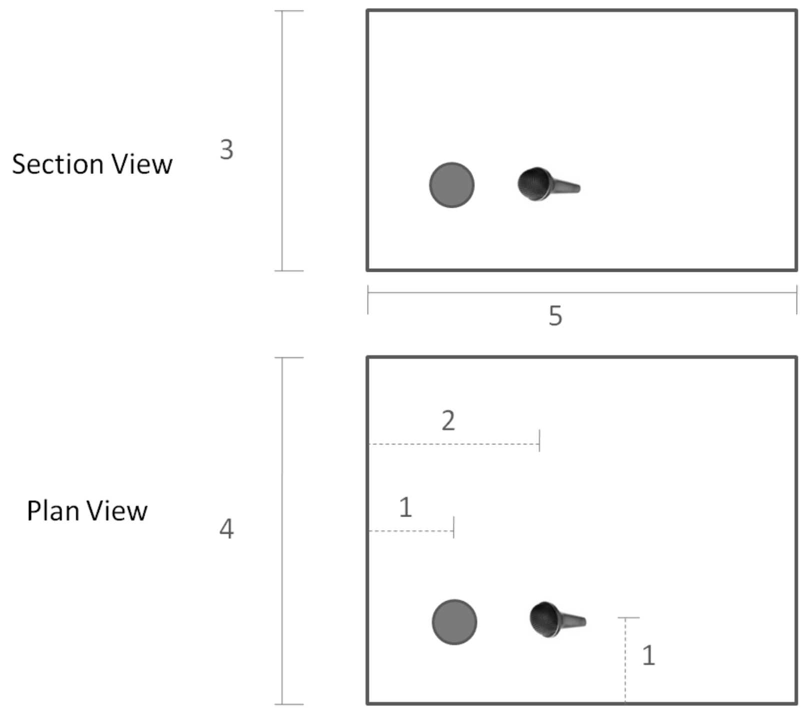


Figure 7: Virtual Room Set Up for Experiment

Section (top) and plan bottom) view of artificial room used in impulse responses for experimentation. Circle represents omni-directional source, microphone represents one channel of a binaural receiver. All measurements are in meters.

Although arbitrarily chosen, these parameters were thought to be fairly reflective of the dimensions of a midsized conference room, a fairly typical speech recognition environment. The source-to-receiver distance was set to a whole number as a practical matter (a coding preference). Although it might admittedly have been a bit smaller for a typical ASR setup, since the main limitations of reverberant ASR will result from having these larger distances this is perhaps not as much of a stretch.

4.1.5 Black Box Experiment

Contrary to many speech recognition experiments focused on improving accuracy of the output, this experiment treats the speech recognizer as a constant (the system under test) and varies inputs to identify its characteristics. The inner workings of the system are not called into question or attempted to be improved upon. In this way the speech recognizer is treated as a black box, and the system identification task at hand is to show the effect of reverberation on its recognition accuracy.

For the preliminary test the average absorption of all materials in the room were varied from 0% to 100% in increments of 5%. (The 100% absorption case was actually specified as 99.99% absorption, to prevent a null signal convolution for the tests on only the reverberant tail. This data point is unrealistic and had some idiosyncrasies in its modelling, but is included here for completeness and ease of reporting. See Section 4.2.3 for further discussion.) Disregarding air absorption, the 0% absorption case would hypothetically result in an infinite signal. However the mirror room algorithm applies a flat 2% transmission loss to all simulations, thus skewing the Sabine/Eyring predictions somewhat. The reverberation times for each average absorption are listed in Table 3, measured as T30 (an extrapolation of the best linear fit of the first 30 dB of decay to the 60 dB drop time). Both the average absorption and the reverberation times reported represent values which were set across all octave bands. Although a flat reverberation spectrum is an unrealistic room setup, having the experimental control of a flat spectrum will result in stronger conclusions about the overall absorption properties. By removing the frequency-dependent component of the analysis for this study, the frequency effects on ASR can easily be compared in future studies. Data for clarity, definition, and speech transmission index are also listed. The room acoustics parameters (other than STI) were processed using a Matlab toolbox developed by the author in conjunction with ARCH 6870 Sonics Research Lab, led by Ning Xiang at Rensselaer Polytechnic Institute. The STI calculation was performed with LexSTI, a freeware STI calculator developed by the Research Division of the Lexington Center and School for the Deaf [43].

Table 3: Absorption Coefficients and Room Acoustic Parameters

Abs. (%)	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
RT (s)	2.08	1.04	.69	.52	.41	.35	.30	.26	.23	.21	.20	.18	.17	.16	.15	.14	.14	.13	.13	.13	.13
D/R	1.8	2.1	2.5	2.9	3.4	4.5	4.4	5.6	5.1	6.9	8.7	10.2	8.7	11.8	14.6	15.3	22.5	17.5	23.2	24.9	28.5
C50 (dB)	-1.7	1.9	4.5	6.4	8.7	9.4	11.0	12.5	14.3	15.2	16.8	18.7	20.5	21.7	23.9	26.0	27.5	29.9	31.6	33.8	33.9
C80 (dB)	-4.4	-1.5	0.5	1.7	3.0	3.1	3.7	4.0	5.2	5.9	6.7	8.0	8.7	9.7	10.1	11.5	12.6	13.9	15.0	16.6	16.7
D50 (%)	27	42	54	60	67	68	71	73	78	81	84	87	89	91	92	94	95	97	97	98	98
STI	.43	.57	.66	.68	.71	.75	.77	.77	.80	.80	.80	.81	.84	.85	.86	.86	.88	.89	.90	.92	.92

A graphical display of the Speech Transmission Index with generally accepted human speech intelligibility correlates is shown in Figure 8. STI values above 0.6 represent good speech intelligibility for human listeners, with close to 100% accuracy, and STI values below 0.45 represent poor speech intelligibility with approximately 30% accuracy [44]. This suggests that, for this model, a human would have very good accuracy when identifying a word from the choice of a small vocabulary for 10% or greater average absorption.

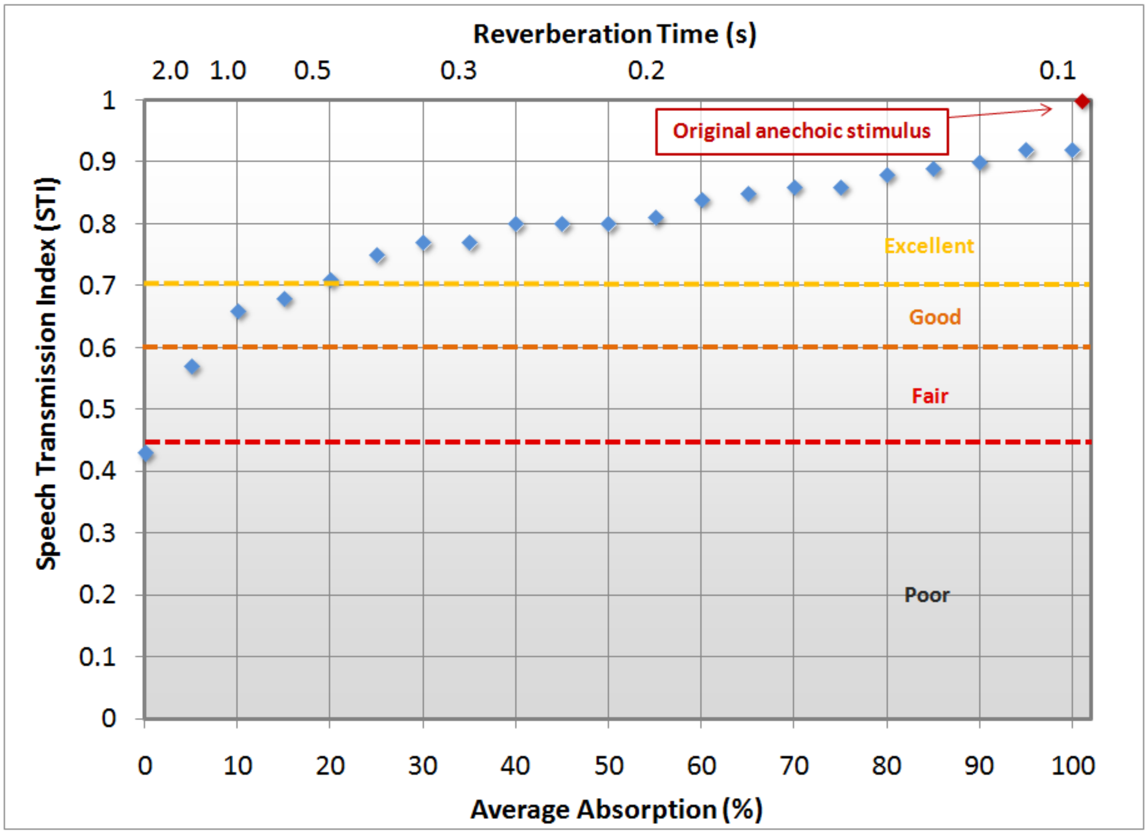


Figure 8: Experimental STI versus Average Absorption and Reverberation Time

Speech Transmission Index calculated and plotted for the transmission paths modeled by impulse responses used in ASR experiment. Original anechoic stimulus plotted at above 100% absorption, all other data points graphed versus their average absorption. Speech quality rankings are specified and separated by dotted lines.

For each of the tests, the default procedure (for the Matlab platform) is to use the exact same file for training and testing. This allows the program to reach 100% accuracy when the two files match completely, which is a good benchmark to start from. As a reality check, the experiment was also run with different file sets forming the training and testing material, similar to the number test described before. The Dragon platform uses HMM prediction models and is therefore using different testing and training material at all times. Additional tests were conducted to analyze the individual effects of early reflections and the reverberant tail on the recognition rate with various configurations of their energy balance.

4.2 Results

4.2.1 Matlab Platform

The results from the primary test of accuracy versus absorption are shown in Figure 9. The solid line is percent correct recognition of all sets averaged together. The error bars represent the standard deviation of all of the sets. (Standard deviation was chosen instead of variance, so that the magnitude could be comparable to the average, rather than an order of magnitude smaller from decimal multiplication). Basic statistics argue that 68% of the population data will be expressed by the full range of one standard deviation above or below the average [5]. The dotted line at 17% accuracy represents 1/6 correct guesses for each set, which is equivalent to random guessing. This represents a baseline for the minimum performance of the recognizer in any trial, as well as the suggested baseline from the ANSI speech intelligibility standard which is a model for this testing. The logic of the Matlab recognizer resulted in minimum performance when its feature vector minimization algorithm found all words to be very far from the training file word, with practically zero similarities. In these cases, it returned the first word in the set as its guess and was therefore right 17% of the time. The results show 100% accuracy for 100% absorption, which is practically anechoic with a reverberation time of 0.1 s (though ideally 100% absorption is by definition anechoic). There is a steady degradation of recognition accuracy with decreasing average absorption percentage starting at approximately 70% absorption. Below 10% absorption there is a sharp drop-off in accuracy from 90% to 60%. At 0%, the accuracy data point intersects the line representing the probability of a correct response by random guessing, which represents minimum possible performance for this algorithm.

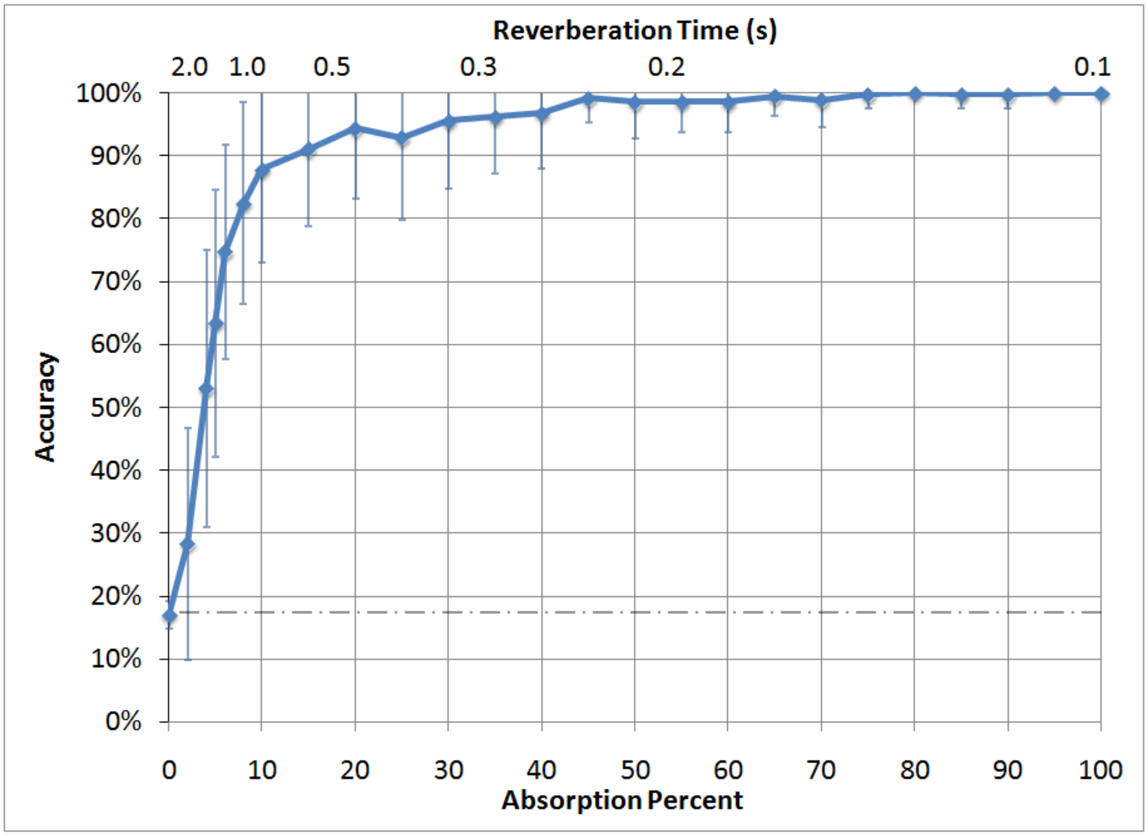


Figure 9: Accuracy versus Absorption for Matlab Recognizer

ASR accuracy is plotted versus average room absorption of impulse response. The dotted line at 17% represents probability of forced choice random guessing, (i.e. the lower limit of performance for this algorithm).

The standard deviation of this measurement and of many subsequent measurements is fairly high. See Appendix C for the individual results of all sets for the Matlab accuracy versus absorption values, as well as reported values for the averages and standard deviations picture above. The standard deviation is less than 5% for only 0% absorption and above 80% absorption. It is greater than 10% for all values from 2% to 30% absorption, and greater than 20% for 10% and 15% absorption. As is borne out by further investigation, there is a fair amount of variation in the degradation pattern of the various word sets versus average absorption. Still, the average shape of the curve shows a significant difference between low absorption and high absorption recognition.

Given that there is so much variation across every data set, it is unsurprising that there is a fair amount of variation when, in Figure 10, the results are separated out into the groups of phonetic similarities. Although the group which only differed in vowel sounds does have a somewhat different trajectory, it should also be noted that this group was the smallest sample set with only 6 sets of self-designed words, since this category was not a part of the ANSI standard.

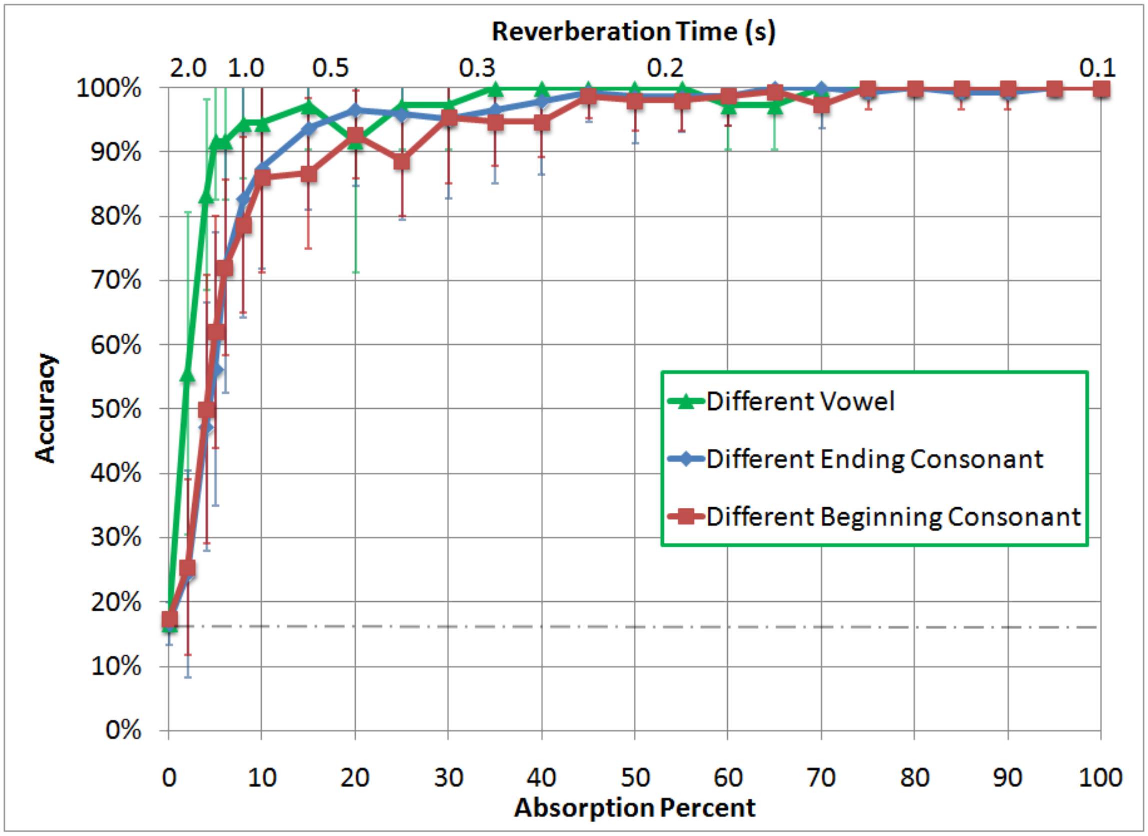


Figure 10: Accuracy versus Absorption Separated By Set for Matlab Recognizer

Shows phonetically separated accuracy versus absorption data; dotted line is probability of random guessing.

The overall results depicted in Figure 9 are based on matching an anechoic training file with several testing samples of the same anechoic file convolved with different impulse responses. To explore whether this change to the basic speech recognition process fundamentally changes the results, a version of the accuracy versus absorption graph was performed with different training and testing files. Four new testing files were created, convolved with impulse responses with an average absorption in increments of 10%, and the accuracy of each set was tabulated and averaged by absorption value. The results are depicted in Figure 11, and reported fully in Appendix D.

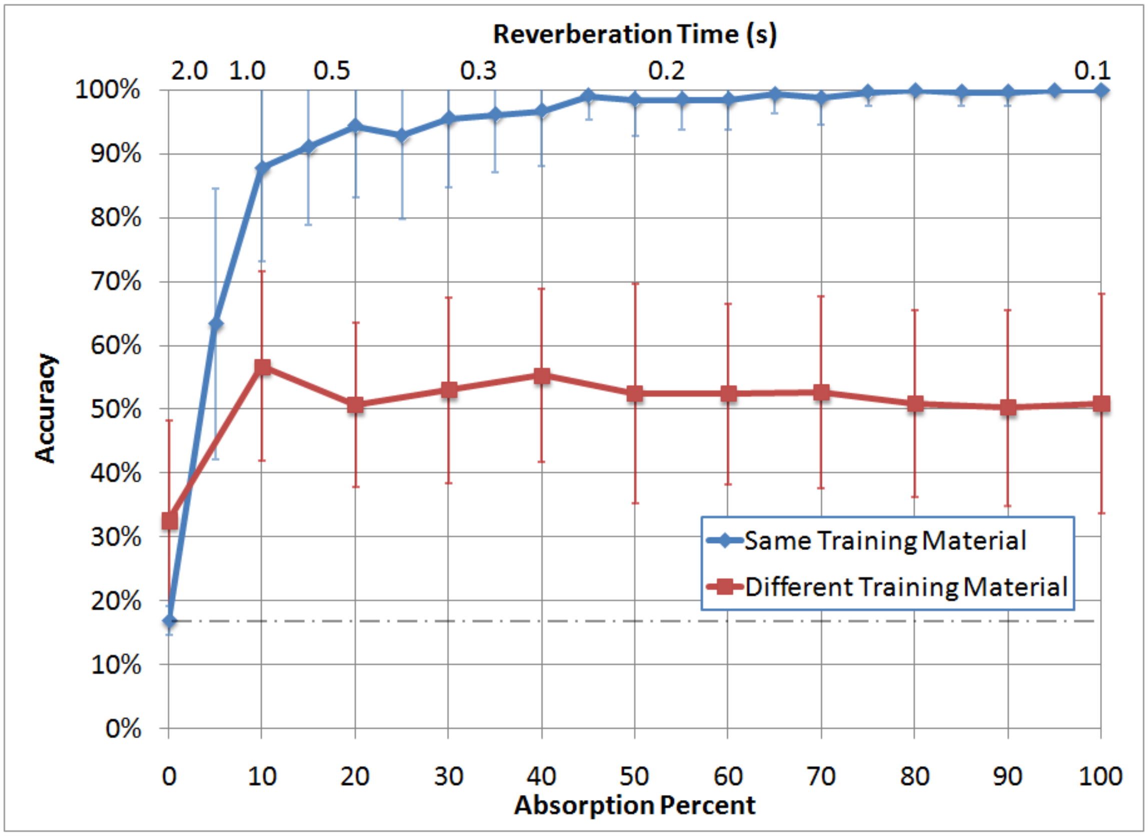


Figure 11: Accuracy Results with Same and Different Training Material

Shows accuracy versus absorption results when different recordings of the training files were tested for recognition (as before, dotted line is probability of random guessing).

The average results are separated into their groupings based on phonetic similarity in Figure 12.

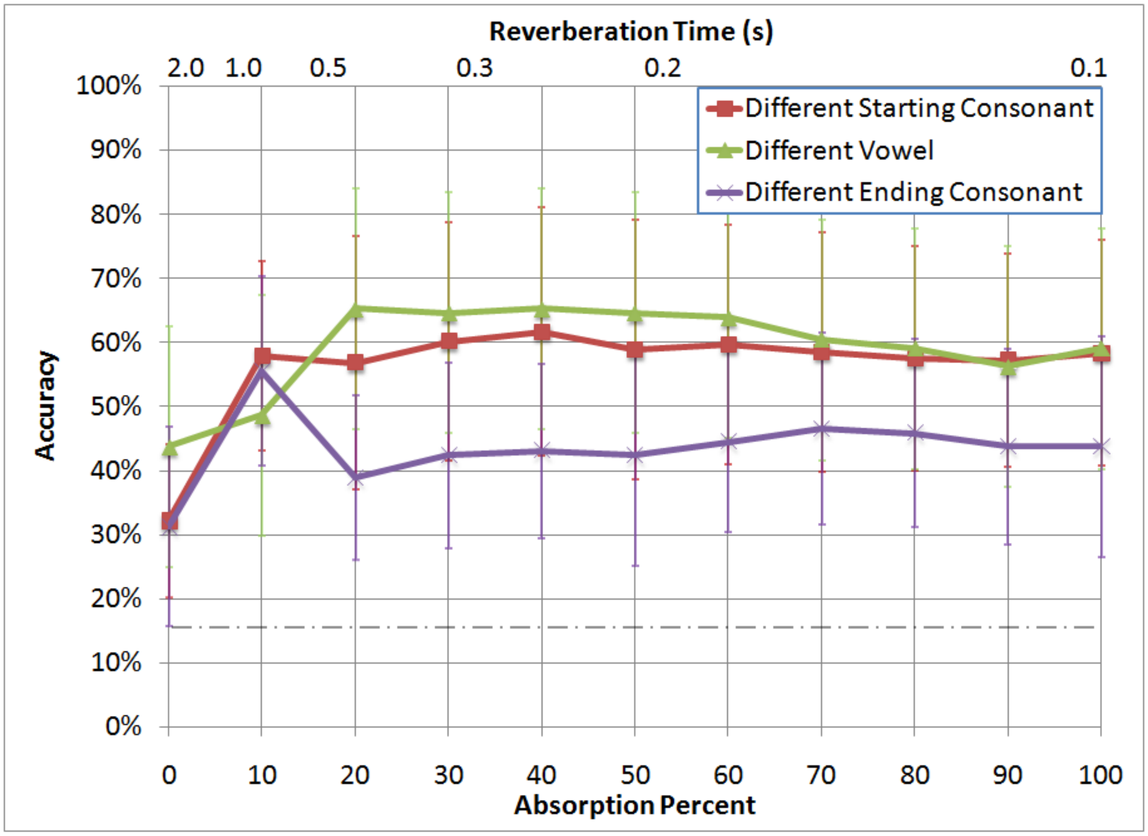


Figure 12: Accuracy Results Different Training Material Separated by Group

Shows accuracy versus absorption results when different recordings of the training files were tested for recognition, separated by phonetic groupings. (As before, dotted line is probability of random guessing.)

The results of the different training material tests show that little new information is added by testing in the more typical setup. In Figure 11, the anechoic accuracy is decreased from 100% to 52% which, while it is fairly typical of a difficult recognition task such as discerning between 6 similar sounding words, limits the range of data that is above the random probability point. It also increases set-to-set variability as represented by error bars, especially as a result of some homonyms which were pronounced differently in one trial versus another. Figure 12 suggests a similar ordering of the three phonetic groups, with different vowel sounds having the best recognition, different beginning consonant sounds having the next best recognition, and different ending consonant sounds having the worst recognition. All of this data is within the limits of the standard deviation of the test, however, and could therefore be attributed to error.

The difficulty in making conclusions with decreased overall recognition and significant additional error, combined with the preliminary test using different training and testing material with the numbers 1 through 10 exhibiting 100% accuracy, gave credence to the choice to base most of the findings of this study on trials with the same testing and training material.

The characteristic of the speech recognition degradation also depends on the contribution from early reflections and the reverberant tail. Using the mirror room Matlab code, the early reflections and reverberant tail are determined separately in two vectors and are subsequently added together. Having tweaked the source code to isolate these effects, several combinations of the degradation on recognition accuracy are shown in Figure 13 to Figure 18.

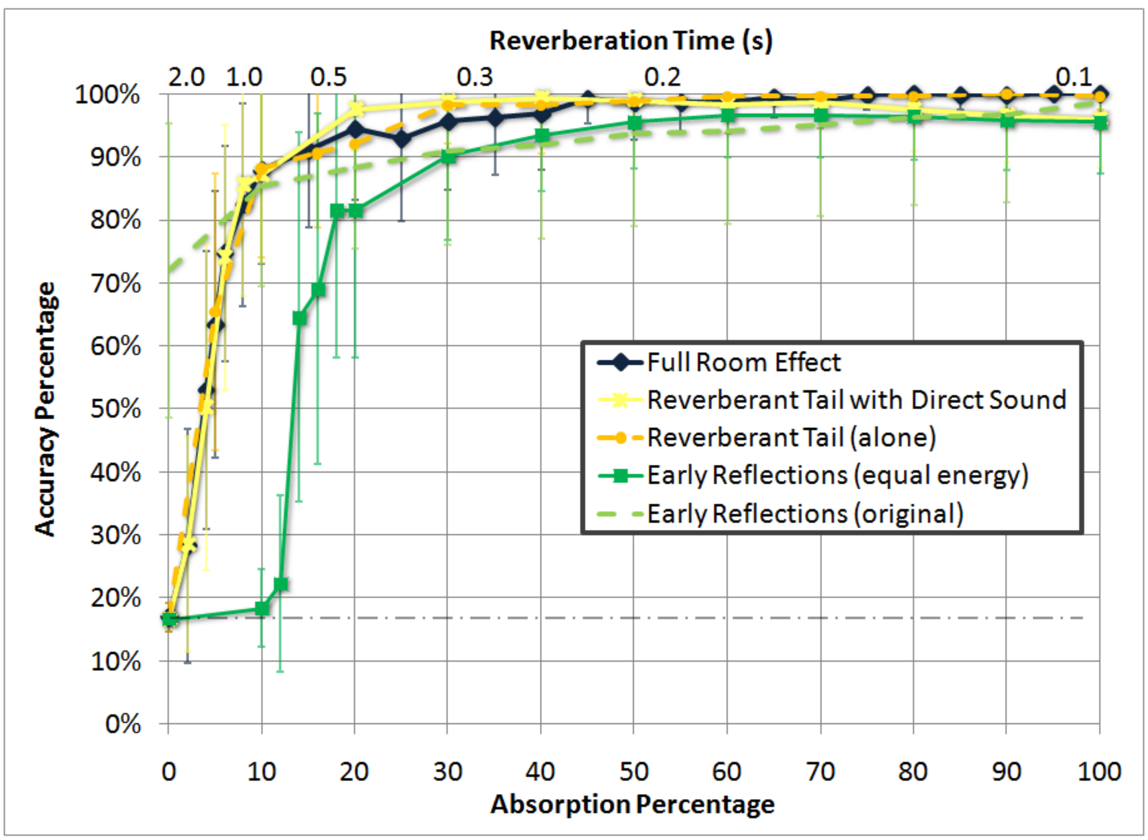


Figure 13: Early Reflections and Reverberant Tail Full Results

Full results of early and late energy exploration, (as before, dotted line is probability of random guessing). See below for additional discussion of separated graphs.

Figure 14 shows an example of the 4 impulse responses compared to the full room impulse response, for an absorption of 20%. The reverberant tail is tested by itself, and normalized to the -1 to 1 range for wave file format. Originally, the reverberant tail did not include the direct sound, since the direct sound is part of the head-related impulse response (HRIR) calculation of the early reflections. A truncation method based on the 0.1 dB point of the Schroeder curve was used to remove just the direct sound portion of the HRIR and add it to the reverberant tail in a separate impulse response. Next just the early reflections were tested. Finally, the early reflections were tested with the spectral power in the early reflections being equal to that in the reverberant tail. A string of zeros was appended to the early reflections to equal the length of the reverberant tail. The root mean square (rms) power of the reverberant tail was divided by the rms power of the early reflections (excluding the direct sound) and that proportion was multiplied by the early reflections. In this case it slightly increased the power of the early reflections relative to the direct sound.

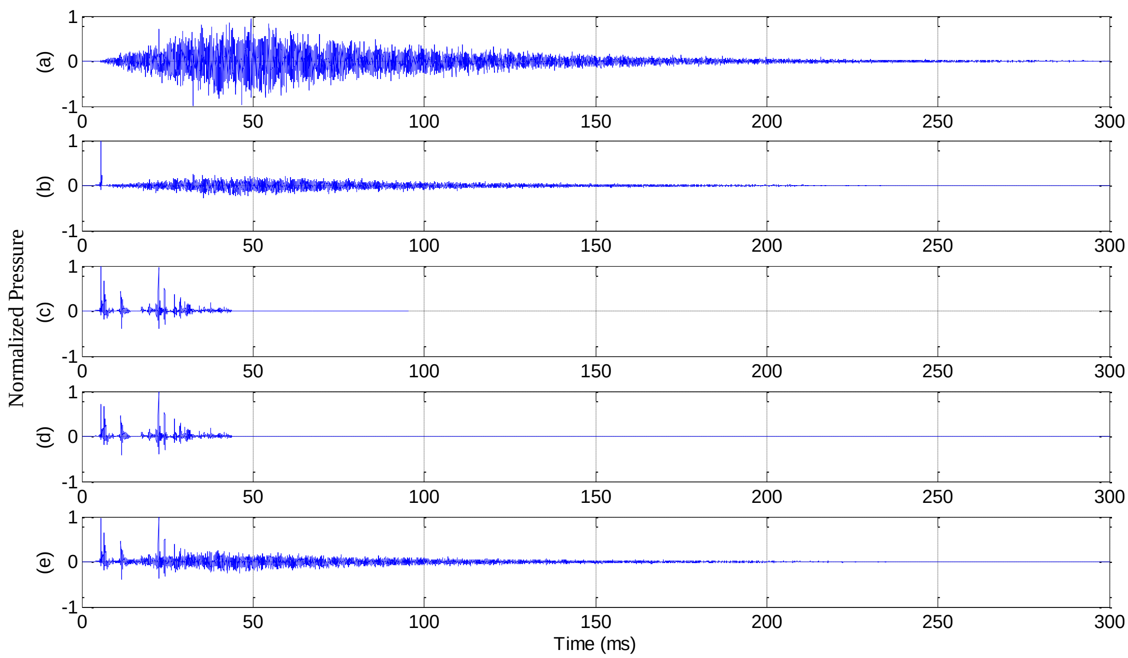


Figure 14: Impulse Response Comparison Absorption 20%

Plotted impulse response components: (a) reverberant tail, (b) direct and reverberant tail, (c) early reflections, (d) early reflections with energy equal to reverberant tail, and (e) full room effect.

Table 4 shows the conversion factors used for normalization of the early reflections to the same power level as the reverberant tail. This normalization added energy to the early reflections up to 30% absorption, and subtracted energy above 30% absorption.

Table 4: Energy Balance

Absorption (%)	0	10	12	14	16	18	20	30	40	50	60	70	80	90	100
RMS_{late}/RMS_{early}	1.98	1.69	1.62	1.56	1.50	1.43	1.36	1.14	0.93	0.73	0.59	0.44	0.34	0.26	0.23

All of the room acoustics parameters which were discussed previously were calculated for the 4 new impulse response conditions at all absorptions. Reported below are reverberation time, clarity (C50) and definition (D50).

Table 5: Absorption Coefficients and Room Acoustic Parameters

	IR	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
RT (s)	Rev	2.08	1.04	.69	.53	.42	.35	.30	.26	.23	.21	.19	.17	.16	.15	.14	.13	.12	.12	.11	.10	.10
	R/D	2.08	1.04	.70	.52	.42	.35	.30	.26	.23	.21	.20	.20	.19	.19	.18	.18	.18	.17	.18	.18	.18
	Ear	.02	.05	.06	.05	.05	.05	.05	.05	.05	.04	.04	.04	.04	.03	.03	.03	.03	.03	.02	.01	.00
	Eq E	.02	.03	.05	.05	.05	.05	.05	.05	.05	.05	.06	.06	.06	.06	.06	.05	.06	.08	.10	.10	.10
C50 (dB)	Rev	-4.1	-.6	2.3	4.6	6.3	8.0	9.9	10.6	12.4	14.3	15.2	16.8	17.2	19.0	20.7	21.6	23.0	25.1	26.3	26.7	27.4
	R/D	-7.6	-4.4	-2.2	-.3	1.3	3.2	4.1	5.5	7.1	8.1	9.9	11.4	12.6	14.1	16.3	17.4	19.3	20.8	22.3	23.7	23.7
	Ear	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
	Eq E	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
D (%)	Rev	28	47	63	75	82	87	91	92	95	97	97	98	98	99	99	99	100	100	100	100	100
	R/D	15	27	38	49	58	68	73	79	85	88	91	94	95	97	98	98	99	99	99	100	100
	Ear	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Eq E	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

The room acoustics parameters reported are specified in the 1st column for each of the 4 impulse responses in the 2nd column: reverberant tail only (Rev), reverberant tail with direct sound (R/D), early reflections only (Ear),and early reflections equalized to energy level of reverberant tail (Eq E).

Since Figure 13 showing all results is somewhat crowded it has been broken up for individual discussions of its components' implications. Figure 15 is a comparison of early reflections to the full room effect containing both early reflections and the late reverberant tail.

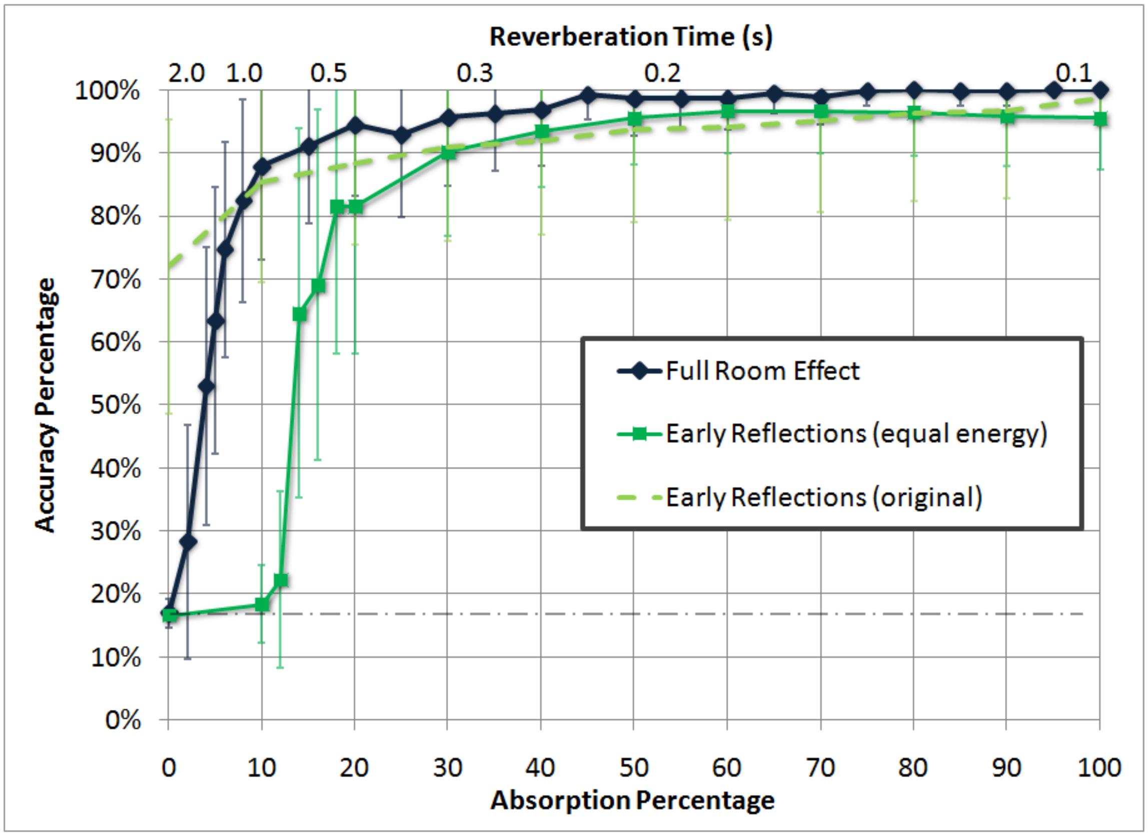


Figure 15: Early Reflections Effect

Comparison of original early reflections, early reflections when equalized to the energy of the reverberant tail, and full room effect.

Without the reverberant tail, the early reflections alone have a better recognition rate than the full room effect for extremely low absorption, and a worse recognition rate than the full room effect for the middle range between 30% and 90%. When normalized for equal energy, the early reflections have a quicker depreciation in accuracy than the full room effect. All early reflections are before 50 ms and based on the 100% definition values, room acoustics theory for human perception would suggest that increasing the direct sound would strengthen the direct sound to provide very clear speech. Here, we see the ASR system is affected more by discrete reflections than a human listener.

In Figure 16, the full room effect is compared to two versions of the reverberant tail vector, with and without the direct sound.

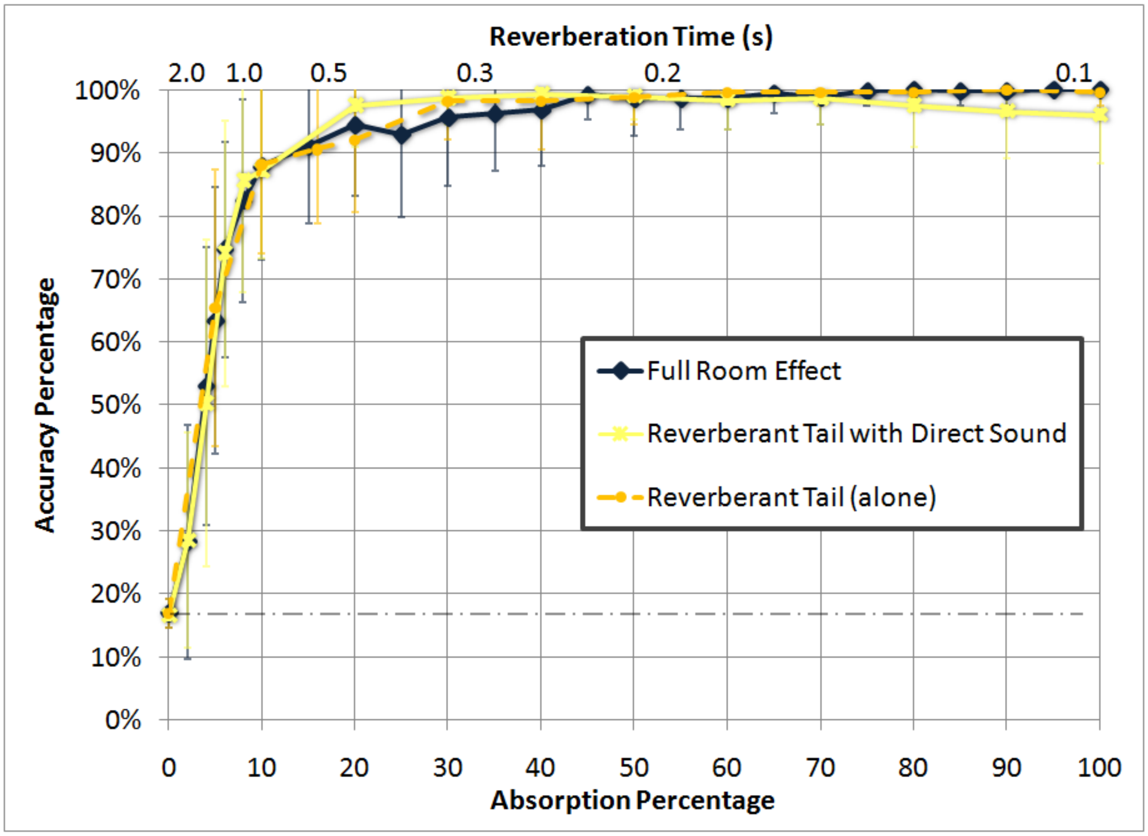


Figure 16: Reverberant Tail Comparison

Comparison of reverberant tail alone, reverberant tail with direct sound added, and full room effect.

The reverberant tail result well mirrors the result of the full room effect, suggesting that late reverberant energy is the primary factor affecting recognition in the full room case. The reverberant tail and direct sound results behave very similarly to the full room effect at low absorptions. At high absorptions its accuracy has a noticeable dip away from the asymptotic 100% of the full room effect, which is probably due to some remnant early reflections from truncation or a strange effect caused by a very sparse, low absorption impulse response. The direct sound clearly provides some additional strength at low absorptions, since the reverberant tail alone has a sharper drop down to the random accuracy level than both the reverberant tail plus direct sound and the full room effect.

Another interesting comparison is the effect of the original vectors of the early reflection and late reverberant tail vectors which comprise the full room effect (see Figure

17). Below 10% absorption the early reflections have a greater accuracy than the full room and reverberant tail vectors. Above 10% absorption the full room effect is close to an average of the accuracy of early reflections and reverberant tail vectors, with the early reflections dropping in accuracy fairly linearly down from 100% absorption.

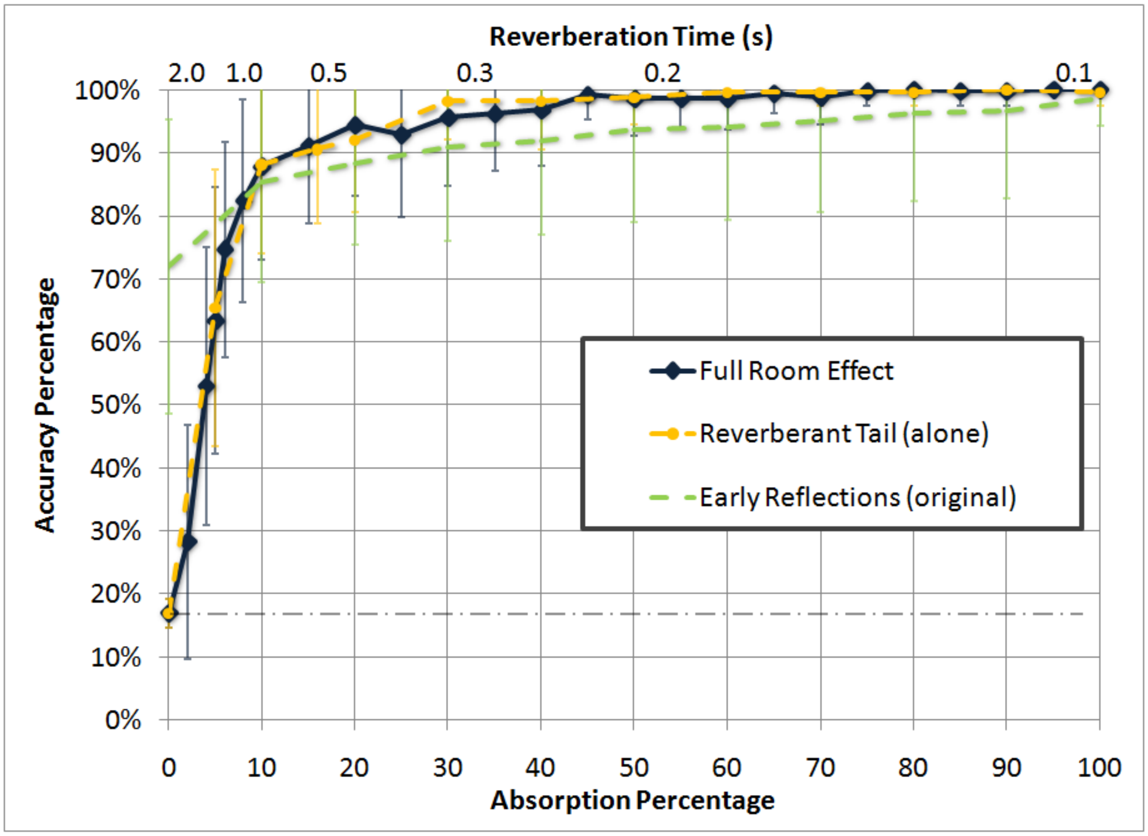


Figure 17: Original Vector Comparison

Comparison of original early reflection accuracy and reverberant tail (no direct sound) with full room effect.

The final comparison (Figure 18) is of the early reflections with energy balance to the reverberant tail with added direct sound, thus it is a comparison of the two effects overall without the energy discrepancy playing a factor.

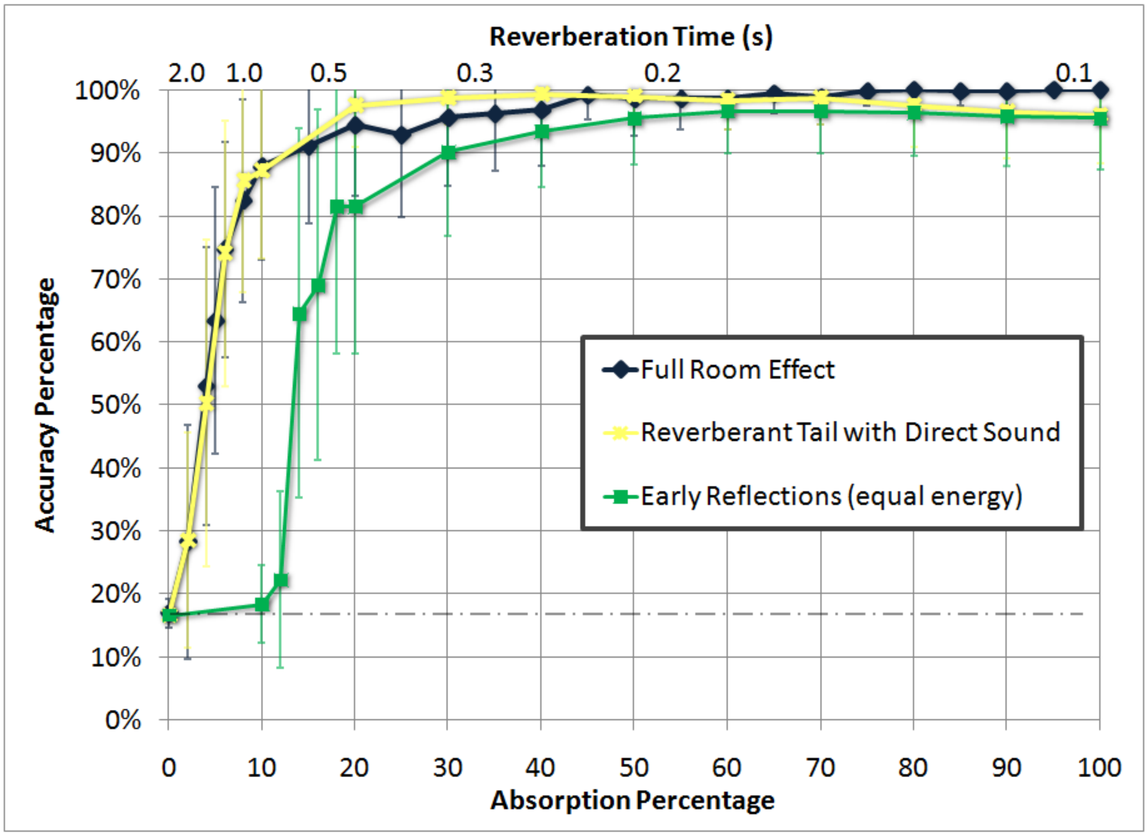


Figure 18: Balanced Energy Comparison

Comparison of reverberant tail with direct sound to early reflections equalized to energy in reverberant tail.

The results show that the discrete early reflections are more detrimental to recognition than a smoothly decaying reverberant tail in all cases. The difference in recognition accuracy at low absorptions is especially large. This is precisely contrary to the measurement of definition at this point, with the early reflections vector having a definition of 100% and the reverberant/direct vector having a definition of 15–30%.

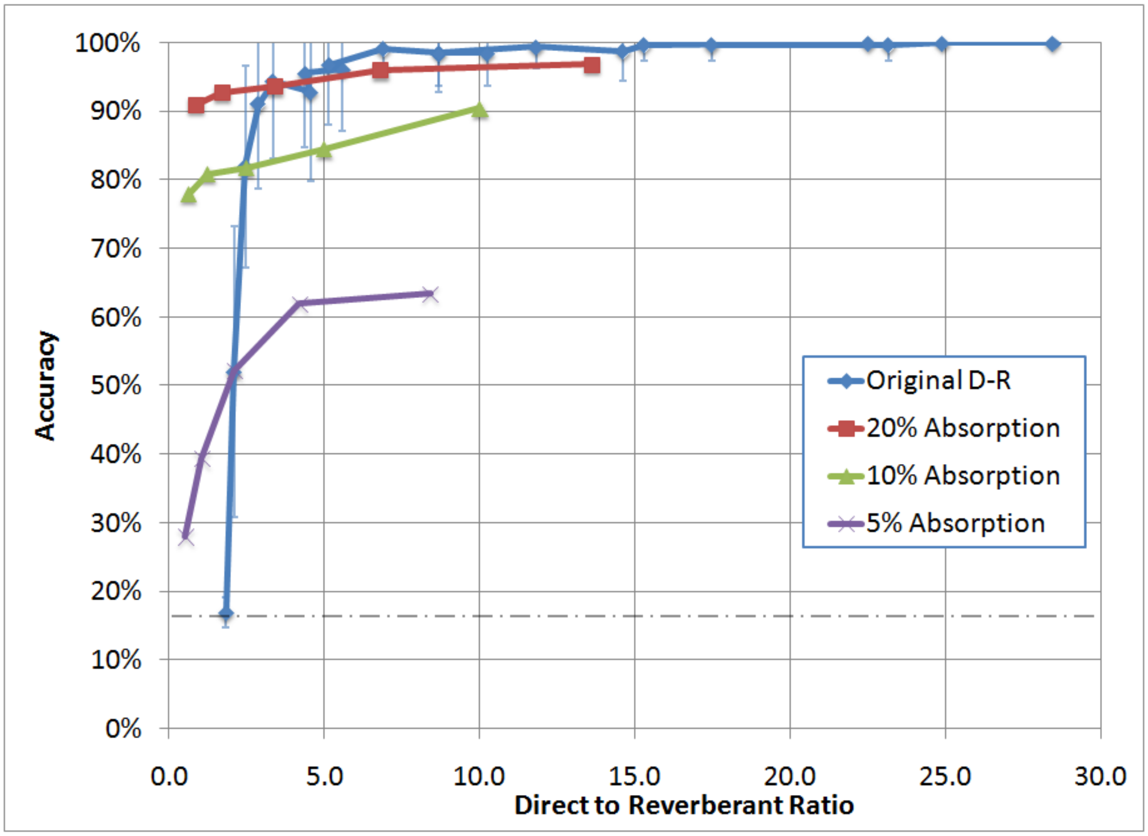


Figure 19: Accuracy Data versus Direct-to-reverberant Ratio

Accuracy Data has been graphed versus its Direct-to-reverberant Ratio. This is compared to the accuracy for D/R altered at three different absorption values.

4.2.2 Discussion of Matlab Platform Results

One of the interesting findings here is the difference between Figure 8, the speech transmission index, and Figure 9, ASR recognition accuracy, graphed on the same axes of absorption percentage in increments of 5%. The shape of these two graphs is similar with a positive overall slope with increases in absorption. The speech transmission graph is much closer to linear, with a slight additional drop-off at 0% absorption (which is a substantially larger reverberation time jump, 1.0 to 2.0 seconds compared to 0.1 second jumps at less absorption). The lower relative position of the speech transmission graph is a result of a 0.7 and higher speech transmission index corresponding to an "excellent" transmission quality, with close to 100% speech identification accuracy. Likewise, the 0% speech identification accuracy point on the speech recognition graph is actually at 17% random probability of guessing and not at 0%. With these two caveats, the two results are surprisingly corroborative. The results for 10% absorption and higher do indeed look like "good" to "excellent" (as assigned in the STI graph) recognition rates, as seen in the ASR analysis ranging from 85% - 100% accuracy. Likewise the 5% absorption point speech transmission is "fair" with 65% accuracy, and the 0% absorption is "poor" with 17% accuracy, the probability of random guessing. One observable difference between the results however, is that there is little middle ground in any of the recognition testing. Again, without knowing the real performance of humans in the same testing scenario, it still seems that humans would have a bit more of a transition period, where they gradually miss more word identifications with decreased absorption but answer others correctly, hovering between 50 and 90%. This is a subject for further scrutiny.

Without substantial and carefully controlled subjective testing, it is hard to address the relationship between this algorithm and human perception with much more specificity. Still, it would seem preliminarily that this algorithm is comparable to the accuracy of human perception in its recognition performance subject to reverberation. Notably, this recognition program is far from a real-time HMM large vocabulary analyzer—it is a discrete word recognizer with a 6 word vocabulary choice, a much longer than real-time processing time, and an algorithm which performs a comparison of the vector distance between two files (which are impulse response convolved versions of the same file). The forced choice word test is harder than other small vocabulary tests because of their phonetic similarities. It is also a realistic scenario, a test used in evaluating the quality of speech transmission pathways. This word recognition algorithm is more accurate and substantially more robust to reverberation than other recognizers.

Another interesting result is Figure 10, the accuracy versus absorption result separated out by their categorical phonetic similarities. It seems intuitive that the sets of 6 words with different vowel sounds would be easier to distinguish than either of the other two sets, with different ending consonants or different beginning consonants. The different vowel sounds (chosen approximately at random) are more noticeably different than some similar sounding different consonants, and last comparably longer than the consonant sounds, presenting the comparison algorithms with much more data. The specific shape of the different vowel curve (with a dip downward for 15% - 25% absorption) is probably due to random variation. With a population of only 6 sets (compared to 25 each in the other two categories), perhaps the different vowel results are showing a slight natural variation which would be clearer in the presence of more data. Other than the substantially better performance of the different vowel set in the 5% - 25% range, the curves of the three data subsets are similar, near 100% accuracy above 50% absorption, and deteriorate quickly to random guessing accuracy at 0% absorption.

The conclusion that emerges out of Figure 11 is, generally, that the performance of the recognition algorithm is noticeably lower and has a smaller y-axis range when trained with different materials than the trials with the same training material. The separation into phonetic groups in Figure 12 shows a somewhat intuitive pattern in the respective accuracies of the three groups (different vowel sounds are the most accurate, different starting consonants are next most accurate, different ending consonants are the least accurate), their differences are well within the margins of errors defined by variation one standard deviation above or below the average. The lower overall accuracy of these results suggests that this task is dominated by the differences between testing and training files. This is due to slight differences in recording technique and in the human subject's speaking of the words from trial to trial, and perhaps larger variations in the way some of the homonyms in the vocabulary were pronounced in various trials. Again, this suggests that the overall properties of this algorithm as a speech recognizer in reverberation are for the most part borne out from the trials with comparisons between convolved versions of the same testing and training file, and the intricacies of the various relationships are better observed from this configuration than in its operation as a true recognizer, by taking out this additional variable.

Finally, an examination of Figure 13 to Figure 18 points to additional insight into the roots of recognition problems in reverberation, whether they are caused by early reflections or late energy. As was observed, both room impulse response components have substantial contributions to recognition deprecation. In isolation, the reverberant energy and the reverberant energy combined with the direct sound (see Figure 16) shows a very similar accuracy depreciation curve to the full room effect, suggesting that the full room effect depreciation is dominated by the reverberant tail. In isolation, the effect of only early reflections (see Figure 15) is still a positive trend with absorption, but with a more gradual slope. The equalization of the energies in the reverberant tail produces a sharper depreciation at low absorptions than either the reverberant tail alone or the full room effect. This is in contrast to room acoustics estimates that all early reflection energy before 50 ms will likely result in high speech intelligibility. Instead, the ASR algorithm seems very sensitive to discrete reflections of large amplitude even if they come relatively early. In examining the original vector comparison (Figure 17) the reverberant tail closely mirrors the full room effect, while the original early reflections vector shows the same general trend with a flatter slope with respect to absorption, producing less accuracy depreciation at the worst points, and more depreciation at the best points. At some points, overlapping of the early and late energy seem to increase recognition, most likely by simultaneously combining the positive intelligibility effects of the smooth energy decay of the reverberant tail and strong early reflections to strengthen the direct sound. It is unclear why this effect would be mainly observed in the region of 10% absorption (with about 1 second reverberation time). The results of Figure 18 show a similar overall comparison between the energy balanced early reflections and late reverberation. There is a slight peculiarity of a drop in accuracy in the reverberant tail with direct sound data at high absorptions (such that it just meets the early reflections graph at 100% absorption), but since the reverberant tail at 100% absorption is an unrealistic acoustical concept anyway, this feature is not especially informative. Comparison of these curves is somewhat more reflective of the true effect of early and late reflections as their root-mean-square energies have been balanced. The two curves show a consistent advantage for reverberant energy, suggesting that strong discrete reflections are more detrimental to ASR than a strong reverberant energy decay.

4.2.3 Dragon Naturally Speaking

The results of the Dragon Naturally Speaking accuracy versus absorption test are shown in Figure 20.

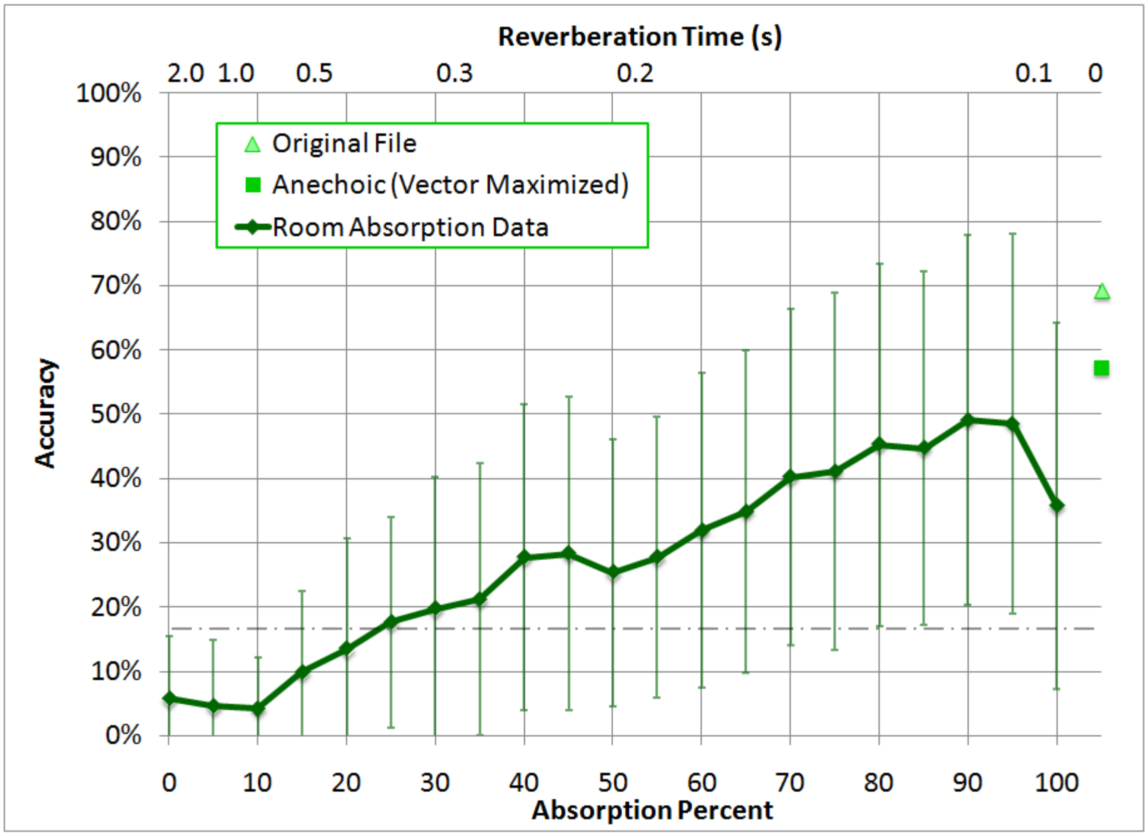


Figure 20: Dragon Accuracy versus Absorption

Dotted line represents chance guessing. Triangle point is original file accuracy; square point is accuracy when vector maximized for -1 to +1 .wav file conversion; connected diamonds are impulse response convolved data points.

The results show an expected reduction in accuracy with decreased absorption. The dotted line representing chance guessing is only included as a point of reference to the Matlab results. It should be noted however that the Matlab program always chose at least one word as its answer (and if it was unable to select one, would return the first word in the set) whereas the Dragon program is much more likely to return no word if confused. The area below 17% is therefore still useful to observe in this case. Notably, the original test file has approximately 70% accuracy—a fair recognition rate considering the atypical usage of individual words for test files. There is a 10% drop in recognition when the file is scaled to the -1 to 1 vector size and saved again as a .wav file, since this represents a difference between the testing material and the overall training material used. There is also a noticeable drop when the 100% absorption impulse response is convolved with the test files; however the salience of this result is limited by the peculiar definition of absorption by the Matlab model. Shown in Figure 21, the error from the additional reflections at 100% absorption is very large. Ideally there would be no frequency effect of the direct sound only, since the frequency content of a pure impulse is a flat line. Figure 21 (bottom) shows the deviation from a flat frequency response. Thus the 100% absorption data should basically be disregarded, since it is far from the ideal representation of a room with 100% absorption.

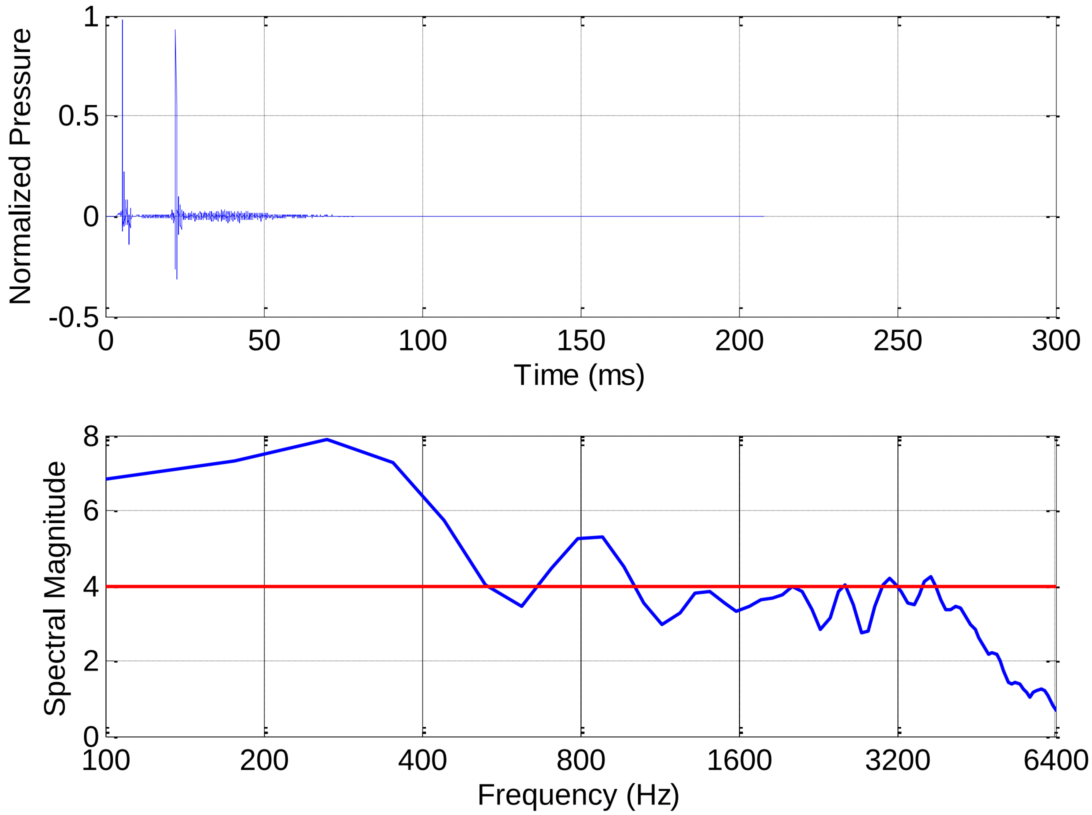


Figure 21: Comparison of 100% Absorption to Ideal Response

The time plot of the 100% absorption case (top) shows a large reflection at 22ms which dramatically affects the frequency response. The ideal spectral magnitude (bottom) would be a flat line (in red), however the measured spectrum is significantly altered (blue).

Although 100% absorption would typically result in only a direct sound, this model adds a very small amount of early reflections and reverberant tail to the direct sound. For 95-100% absorption this result shows a dramatic decrease in the accuracy of the Dragon HMM recognizer, pointing to a possible sensitivity to slight time domain blurring or frequency spectrum comb filtering. In general this model is fairly linear, suggesting something close to a 5% drop in accuracy per 10% drop in absorption. There is a large variability in the results, due to a poor recognition rate for the original and anechoic files. Several data sets exhibited 0% recognition at all absorptions (see Appendix E for all datasets, plus reported averages and variances).

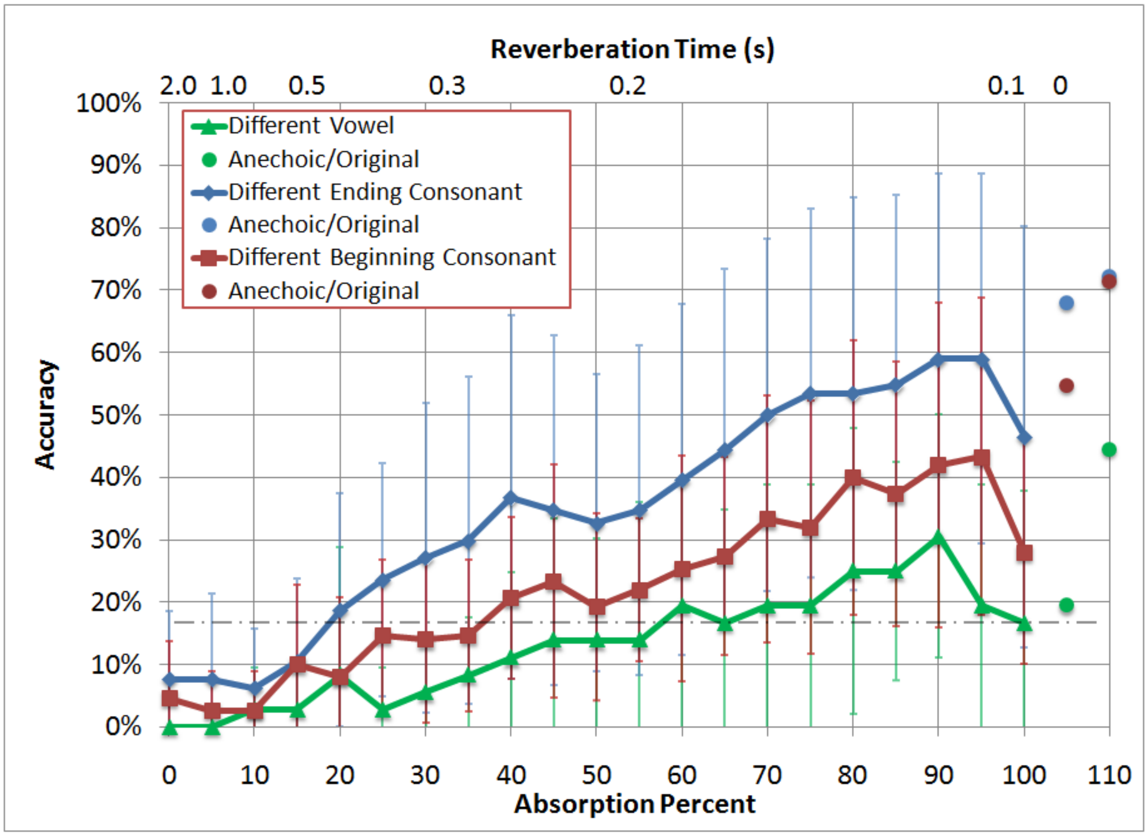


Figure 22: Dragon Separated Data Accuracy Results

Average accuracy reported for each phoneme similarity group. Additional data points for original files (plotted at 110% absorption) and anechoic scaled files (plotted at 105% absorption).

As in the previous section, the results of the absorption test were separated out into their sets based on phonetic similarities. The results of this separation are shown in Figure 22 and reported in Appendix E.

Despite large variability, there seem to be significant differences in Dragon's accuracy levels of the phonetic groups. The results seem to suggest that the recognition rate of words with different consonant endings have the best recognition, words with different beginning consonants have slightly poorer recognition, and words with different vowel sounds have the worst recognition. That ranking, however, is counter-intuitive to what one might expect about reverberant speech recognition, and in conflict with the results of the Matlab test. A confounding factor, which is not explored here, is the vocabulary models Dragon uses for its recognition. Some of the words, especially in sets with different vowel sounds, were nonsense words spelled phonetically, for example, Set 52: grin, gran, groan, gren, green, and groon. The Matlab program is a feature vector comparison and did not take into account whether the words were English words or not, whereas the Dragon program uses built-in phonetic models of the words in its vocabulary to perform the recognition. Without a model for the nonsense words, its recognition will be based on the spelling of the word, but will likely be poor compared to real English words for which the HMM has several speech models. In the same vein, the actual rates of accuracy degradation are similar and actually slightly steeper for the words with different ending consonants, since its recognition is much higher at high absorption and approximately the same at low absorption. In turn, a relative slope argument maybe confounded by the leveling off of all three datasets at 0% absorption, since 0% recognition is the global limit of the tests. To help address which of these factors is at stake, the data sets were graphed with only 13 sets averaged in which all start from 100% recognition with original files. There are only 8 in the different ending consonants group and 5 in the different starting consonants group, so the variability of these results is very large. There are no sets with 100% original recognition in the similar vowel sounds group.

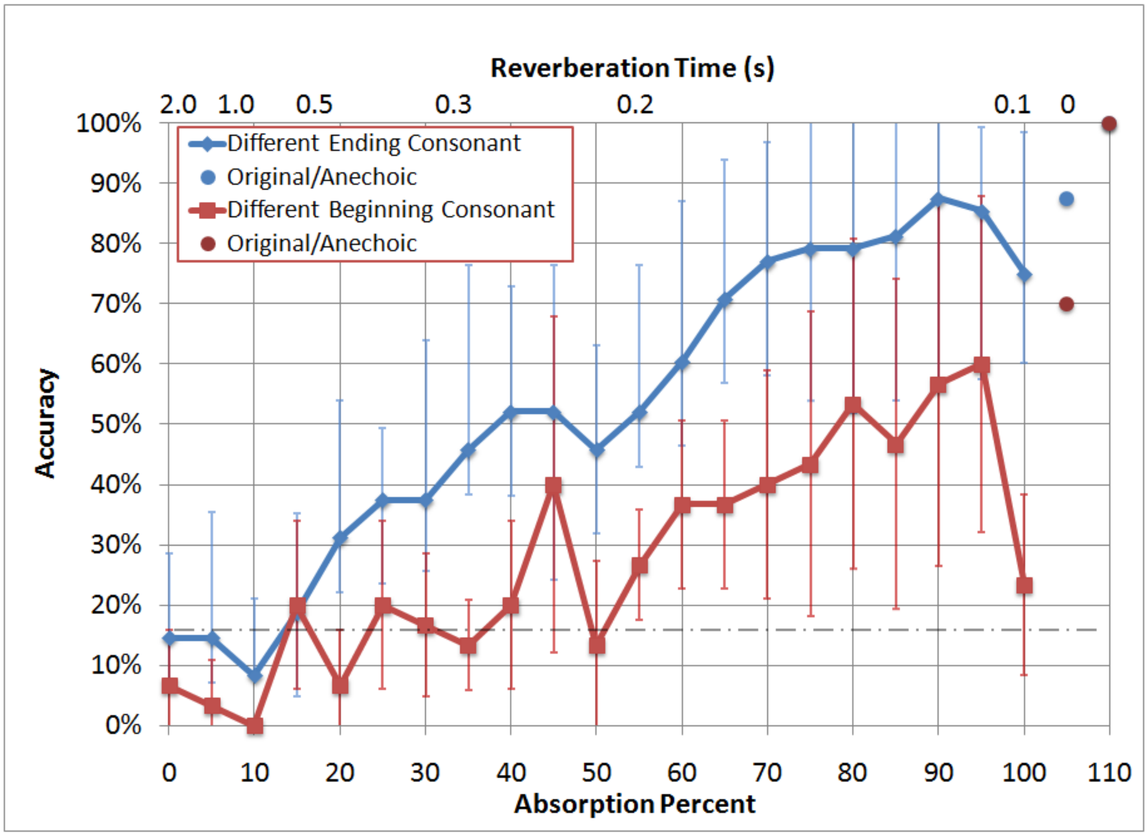


Figure 23: Dragon Separated Data Starting from 100% Accuracy

Separated accuracy versus absorption data, comparing only datasets with 100% original accuracy. The similar vowel sounds cannot be graphed since there are no datasets with 100% original accuracy.

While the decrease in the number of samples has increased variability widely, the average behavior of these two data sets is telling. The different ending consonant group has a much sharper slope suggesting a large dependence on excess reverberation masking the ending sound. The different starting consonant drops substantially in the initial anechoic scaling and in the low absorption values, suggesting less of a dependence on reverberation and more on a mismatch of feature parameters with the training set.

An additional analysis was conducted of the effect of early reflections and late reverberant energy on Dragon's accuracy, shown in Figure 24.

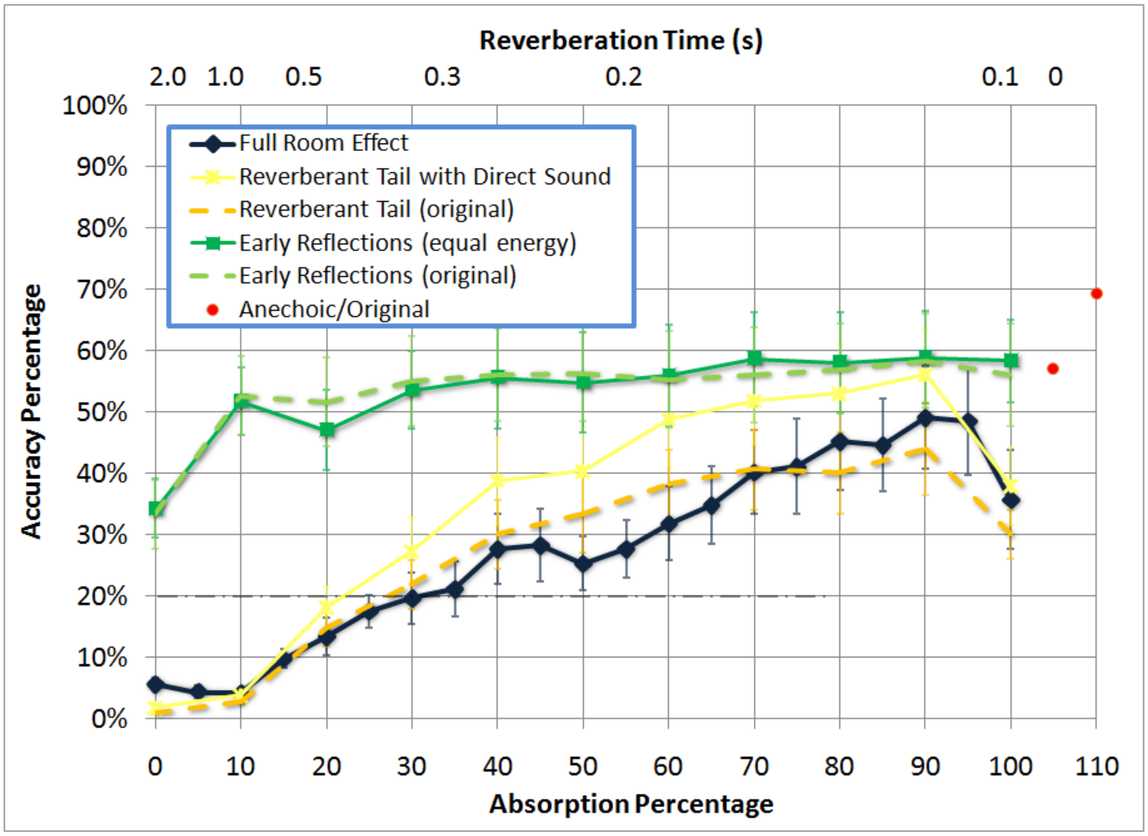


Figure 24: Dragon Early and Late Energy Analysis Full Results

The full results of early reflections and late reverberant tail are compared to the full room result. Original and anechoic but scaled accuracy data points are shown as a point of reference.

The results show almost no depreciation from early reflections at high absorptions, whether weighted with equivalent energy to the reverberant tail or not. The starting point of accuracy for the early reflections aligns closely with the rescaled anechoic vector accuracy, unlike either the Matlab result or the other vectors pictured here. The shape of the depreciating accuracy from early reflections curve is similar to the Matlab result, however.

The accuracy depreciation curve for reverberant energy is similar to the full room effect, as was the case in the Matlab result. The overall accuracy depreciation seems to be dominated by the diffuse reverberation and characterized by the absorption coefficient. The vector with reverberant energy only (no direct sound) is very similar to the full room effect with some points of slightly better and worse accuracy around 10%. The vector with reverberant energy and direct sound actually has a much better accuracy than the full room effect. Although the early energy has little effect on the accuracy alone, it depreciates the accuracy in of the overall room when combined with the reverberant tail. This makes little sense from the perspective of which vector causes more damage, but can perhaps be interpreted as an overall energy balance increase in the energy after the direct sound. If in the full room effect the energy balance is

Direct Sound : Early Reflections + Reverberant Tail

perhaps it also makes sense that this performs worse than either of

Direct Sound : Early Reflections

or

Direct Sound : Reverberant Tail .

In any case it is a different result and would lead to a different explanation than the Matlab word recognizer result.

4.2.4 Discussion of Dragon Results

Before a full comparison of the results will be an analysis of the Dragon results without necessarily extrapolating to all speech recognizers. One major aspect of this research is the need to quantify and qualify the degradation of speech recognition, and that is a problem which should be looked at on an algorithm-specific microscopic level.

The result in Figure 20 shows the overall behavior of accuracy degradation with a loss of average absorption. Although there was significant variability between sets, a general pattern could be seen in the relationship between accuracy and absorption. To compare again the ASR accuracy to human speech intelligibility, this recognizer is significantly worse than a human listener even at very small absorptions. The overall accuracy is low because of the atypical use of the continuous speech recognizer as a word recognizer and because it was not a forced choice test with 1/6 probability of guessing right. The Dragon recognizer would report no word if it was unsure of what it heard. There were also several nonsense words which would confuse the recognizer.

This points to a problem with drawing a significant conclusion from Figure 22 since the differences in subset curves might be based more on the overall recognition rates of the subsets than a difference in the behavior of their response to reverberation. Since only 13 out of 56 sets had perfect recognition with the original test files, it is hard to have a statistically significant evaluation of their performance. The additional representation of this data in Figure 23 corroborates this, as when the average is graphed only by datasets starting from 100% accuracy, the slope of the different ending consonant sound group is sharper and appears more related to absorption than the slope of the different beginning consonant group which seems to be more accuracy degraded by the vector scaling process and at comparably low absorption values.

The early and late energy analysis in Figure 24 shows that the full room effect is dominated by the reverberant tail. The early reflections by themselves degrade accuracy very little and with only slightly increased degradation at lower absorptions. They still have a significant effect in the overall result, since both the reverberant alone and the reverberant/direct vectors have different accuracy curves than the full room effect. The vector with combined reverberant and direct sound has a 10-15% improvement in accuracy over the full room effect at middle absorptions (40-70%) suggesting that early reflections increase the amount of energy after direct sound and degrade recognition.

4.2.5 Comparison of Dragon and Matlab Results

The two recognizers were tested and each performed essentially all of the same recognition experiments. There are significant similarities between the results of the full room effect accuracy versus absorption results. Both recognizers exhibit accuracy reduction with absorption. The Matlab word recognizer is close to completely accurate down to around 30% absorption with 95% accuracy, and then within 30 – 0% absorption change has a sudden drop down to its accuracy floor of 17%. Contrarily, the Dragon recognizer has a steady decline of accuracy starting from 100% absorption. The changes in reverberation time are so slight above 40%, that this somewhat linear relationship is almost surprising. The Dragon recognizer appears to be sensitive to small changes in the strength and duration of reverberant energy as it increases the feature distance between the training set and the testing set. The Matlab recognizer, because it is a feature vector comparison and truly a forced choice tester, is significantly less sensitive to small changes at high absorption. That is, the Dragon recognizer accuracy plummets if its testing file is altered from its vocabulary model and returns no word if it cannot make a match, whereas the Matlab recognizer chooses the closest vocabulary model to its test file making it significantly more robust.

The graphs showing separation by phonetic grouping also show different results. The Matlab result points to an accuracy increase in the different vowel sound words at middle absorption values, and an accuracy decrease in words with different ending consonants. The results were not definitive based on a wide variation. The Dragon results were dominated by the original relative accuracies of the groups. No general conclusions can be drawn with respect to the different vowel sound group, since there were so many nonsense words in each set. When the 100% original accuracy sets were graphed for each of the remaining groups, the Dragon data shows a sharper accuracy decline with reverberation for words ending with a different consonant sound. In general, out of both recognizers, words ending in a different consonant sound may be the least recognizable group when subjected to reverberation.

Finally, the energy analysis provides an interesting contrast. Although the Matlab recognizer shows depreciation of accuracy from early energy alone, the Dragon recognizer shows little further depreciation than was already introduced from rescaling of the .wav file. The depreciation slope with decreased absorption from early reflections alone is similar from both recognizers. When the early energy is normalized to equal energy, the Matlab recognizer shows a sharp decline in accuracy, while the Dragon recognizer has the same accuracy as the original early reflections. The overall room effect accuracy depreciation in both recognizers is based primarily on room absorption in the reverberant tail. The addition of early reflections to the reverberant tail has generally negative effects in both recognizers, with reverberant energy or reverberant energy with direct having a higher accuracy than the full room effect.

4.2.6 Comparison to Prior Work

It is important to frame these results with respect to the prior work in the field. The main study which provided data on ASR performance with variable room acoustics parameters was Palomäki et al. [21]. Their study was also conducted using convolution with artificial room impulse responses. The room size in their study was slightly smaller than this experiment, though comparable—6 by 4 by 3 m (Palomäki) to 4 by 5 by 3 meters.

Since their experiment was so well documented, a direct comparison can be extrapolated to these results. The overall results of this study are graphed versus the Palomäki et al. data in Figure 25.

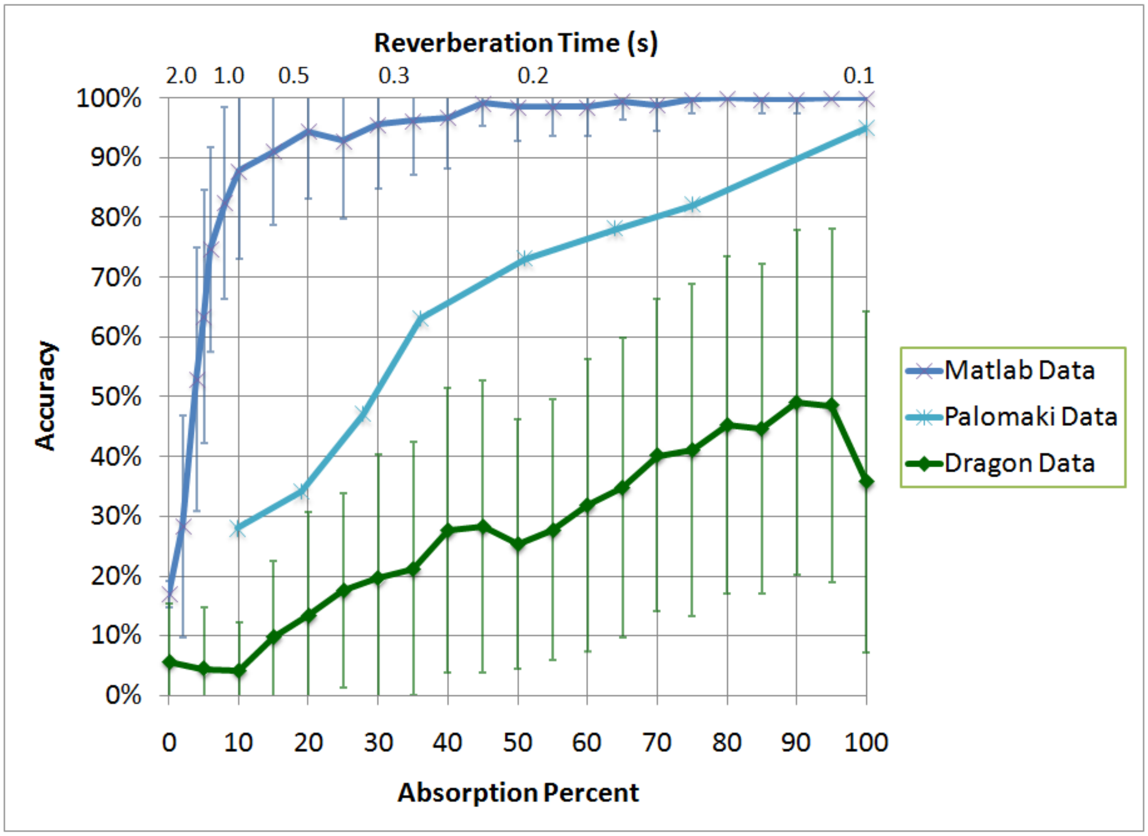


Figure 25: Comparison to Palomäki et al. Data

The Palomäki et al. result shows a similar slope to the Dragon data. The overall position of the data may be slightly skewed because of the room size. Even so, this figure shows three distinct accuracy curves with respect to room absorption. This may point to a difficulty in trying to precisely characterize or address the problem of reverberant ASR. There is a wide variation due to testing procedures and algorithm used.

To account for the discrepancy in room size between the two studies, and in order to include several more studies which listed reverberation time without room absorption values, the results have also been graphed versus reverberation time in Figure 26. Several studies are only one point on the graph.

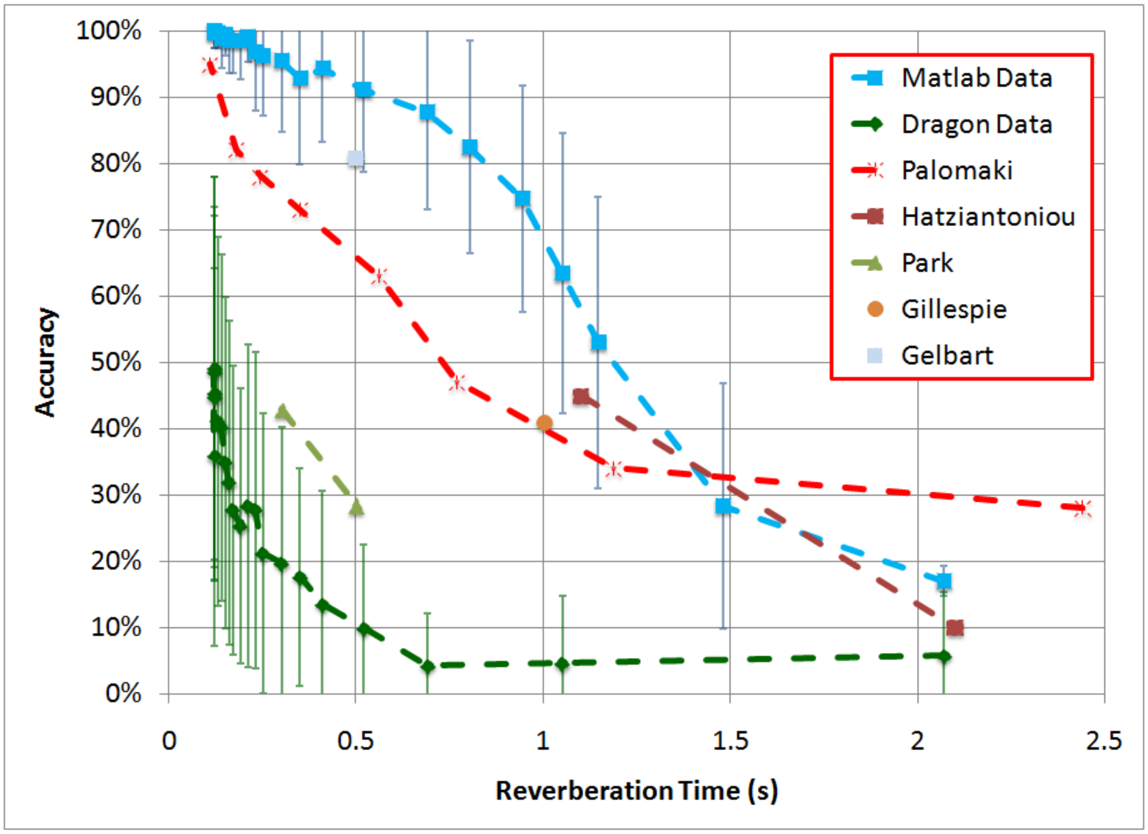


Figure 26: Comparison of Reverberant ASR Studies

This comparison suggests that most of the data found from the platforms in this study are on the extremes of the overall ASR performance in the field. The Matlab recognizer is more robust to reverberation time than most data gathered from other studies, while the Dragon recognizer is less robust. The reason for the large discrepancies in the ease of these tasks was discussed in the previous section. The overall slope of the decline in accuracy with increasing reverberation time is comparable, though the shapes have some differences. The Matlab recognizer appears to have a more negative second derivative than the Palomäki et al. data, for instance, with a general curve down with increasing RT. The Palomäki et al. data is fairly linear with RT at about 35% accuracy drop per additional second of RT, until the last data point at 2.5 s RT, which does not follow this trend. The Dragon data appears to have a more positive second derivative than the Palomäki et al. data, with a generally curving upward shape (though not with a positive slope). The Dragon data is bounded by a zero recognition asymptote, so this shape could be affected by that as well. The Hatziantoniou data points show a similar linear trend to the Palomäki et al. and Matlab curves, with about 45% accuracy depreciation per second of RT though they are in a region where each of the other data sets has few data points. The Park data shows a sharper depreciation between its two data points, about 70% accuracy depreciation per second of RT, though its data points are very close together in a low RT range, so any extrapolation outside of this range is tenuous. Its slope and accuracy position is more similar to the Dragon data than any other data points. Overall, there is a large observed variability between the accuracy degradation with respect to RT, as there are many test methodologies, experimental setups, and algorithms available for ASR evaluation.

5. CONCLUSIONS AND FUTURE WORK

Having first approached the problem of reverberant speech recognition with the goal of implementation improvement, it soon became clear that the ASR field has an abundance of adept computer programmers to explore implementation strategies. Viewed from the science of acoustics, it seemed that rather than some new implementation, what the discussion lacked were experimental data defining the problem. In the literature review, many researchers tested their algorithms in one or two specific reverberant conditions or specified no acoustical parameters of their simulation whatsoever. The goal of this thesis shifted to arrive at a thorough and precise definition of the widely acknowledged but little studied degradation of automatic speech recognition in reverberant sound field.

Two specific speech recognizers were chosen and tested. The Matlab recognizer was chosen largely for ease of use and its flexibility to be reprogrammed for the purposes of specific investigations, while the Dragon analyzer was a commonly used product for typical practical applications. The parameters analyzed were fairly basic: average room absorption across all frequencies, and early or late energy balances in various configurations. The test samples were acquired with a simple anechoic recording process and convolved with a series of impulse responses; the methodology could easily (and should) be repeated with a new system or analyzing new room-acoustics configurations.

The results point to some interesting conclusions for the ASR community. First, in general, the recognition problem seems to be more a masking of consonant sounds by reverberation than a formant alteration affecting the recognition of vowels, and typically masking more late consonants than early consonants. This correlates with human speech intelligibility. Secondly, the degradation seems to be dominated by the reverberant tail (i.e. the late, diffuse-field energy). Although early reflections have an additional effect, the general properties of the room reverberation degradation mirror that of the reverberant tail. When the early reflections are equalized to be as strong as the reverberant energy they sometimes show a substantial accuracy decrease. This does not necessarily correlate with human intelligibility, since early reflections usually strengthen the direct sound and increase intelligibility, and yet they can dramatically decrease ASR accuracy.

Beyond the specific similarities in conclusions based on the two speech recognition algorithms, the many differences in the experimental results of the two systems are important to note as well. In comparison to human speech perception, observations of this system (with the current assumptions and parameters in place) suggest that the Matlab recognition accuracy in reverberant conditions is similar to human speech intelligibility performance, while the Dragon recognition accuracy is significantly lower. As noted previously, the success of the Matlab recognizer does not disprove the problem or break new ground for a speech system which performs like a human. Still, the human-to-speech recognizer performance is a comparison that should be made in the process of such further comparative studies and is an important goal of such a study. The differences in behavior of each recognizer to early and late energy are also significant. The Dragon data shows significantly more robustness to early reflections and direct sound with a reverberant tail than to the full room effect, whereas the Matlab results are close to the opposite. Thus the Dragon deficiencies may be summarized more by the amount of late energy present, whereas the Matlab accuracy depreciation may be based more on the nature of the late energy and the combined shape of the reverberant decay envelope.

The two recognition results were compared to other results in the field to begin to arrive at a consensus on the nature of the depreciation of ASR in reverberation. The results showed a wide variability resulting from testing procedure and algorithm chosen. Nevertheless, they showed some similarities in the slope of the depreciation between systems. Further standardization of testing and reporting procedures with regards the room acoustics experimental setups will help to the ASR field to have a more nuanced and cohesive discussion of reverberant ASR in the future.

Additionally, it should be noted that as one of the first acoustic analyses of automatic speech recognition, the raw data, reported averages, and shape of each accuracy curve represent important findings of the study. They will help inform the recognition community on the nature of the reverberant ASR problem. It is the hope of the researcher that these findings can be a benchmark for further investigation and point to new alternatives and approaches to the problem.

LITERATURE CITED

[1] R. Cole, J. Mariani, H. Uszkoreit et al., Survey of the State of the Art in Human Language Technology, Cambridge, UK: Cambridge University Press, 1997.

[2] B. Plannerer, An Introduction to Speech Recognition, Munich, Germany, 2005.

[3] R. C. Hendriks, R. Heusdens, and J. Jensen, “Adaptive Time Segmentation for Improved Speech Enhancement,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2064-2074, 2006.

[4] L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” in Proceedings of the IEEE, 1989, pp. 257-286.

[5] D. Montgomery, and G. Runger, Applied Statistics and Probability for Engineers, 4th ed.: John Wiley & Sons, Inc., 2007.

[6] N. R. French, and J. C. Steinberg, “Factors Governing the Intelligibility of Speech Sounds,” J. Acoustical Society of America, vol. 19, no. 1, pp. 30, 1946.

[7] R. Thiele, “Richtungsverteilung und Zeitfolge der Schallruckwurfe in Raumen,” Acustica, vol. 3, pp. 291-302, 1953.

[8] W. Reichardt, O. Abdel Alim, and W. Schmidt, “Definition und MeBgrundlage eines objektiven MaBes zur Ermittlung der Grenze zwischen,” Acustica, vol. 32, pp. 126-137, 1975.

[9] J. S. Bradley, R. D. Reich, and S. G. Norcross, “On the combined effects of signal-to-noise level for classrooms from a comparative study of speech,” J. Acoustical Society of America, vol. 107, no. 2, pp. 871-875, 1999.

[10] T. Houtgast, and J. M. and Steeneken, “The modulation fransfer function in room acoustics as a predictor of speech intelligibility,” Acustica, vol. 28, pp. 66-73, 1973.

[11] D. Ruggles, “A Binaural Approach to Speech Intelligibility,” Master's Thesis in Architectural Sciences, Rensselaer Polytechnic Institute, Troy, NY, 2007.

[12] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” J. Acoustical Society of America, vol. 25, no. 5, pp. 975-979, 1953.

[13] N. I. Durlach, “Equalization and cancellation theory of the binaural masking level difference,” J. Acoustical Society of America, vol. 35, pp. 1206-1218, 1963.

[14] M. Lavandier, and J. F. Culling, “Speech segregation in rooms: Effects of reverberation on both target and interferer,” J. Acoustical Society of America, vol. 122, no. 3, pp. 1713-1713, 2007.

[15] J. F. Culling, “Evidence specifically favoring the equalization-cancellation theory of binaural unmasking,” J. Acoustical Society America, vol. 122, no. 5, pp. 2803-2814, 2007.

[16] K.-C. Yen, and Y. Zhao, "Robust Automatic Speech Recognition Using Multi-channel Signal Separation Front-end." pp. 1337 - 1340.

[17] A. Iyer, B. Smolenski, R. Yantorno et al., “Speaker Identification Improvement Using the Usable Speech Concept,” in European Signal Processing Conference, 2005.

[18] K. J. Palomaki, G. J. Brown, and J. P. Barker, “Techniques for handling convolutional distortion with ‘missing data' automatic speech recognition,” Speech Communication, vol. 43, no. 1-2, pp. 123-142, 2004.

[19] M. Holmberg, D. Gelbart, and W. Hemmert, “Automatic Speech Recognition With an Adaptation Model Motivated by Auditory Processing,” IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 1, pp. 43-49, 2006.

[20] G. Brown, J. Barker, and D. Wang, "A Neural Oscillator Sound Separator for Missing Data Speech Recognition." pp. 2907-2912.

[21] K. J. Palomäki, G. J. Brown, and D. Wang, “A Binaural Auditory Model for Missing Data Recognition of Speech in Noise,” Speech Communication, vol. 43, pp. 361-378, 2004.

[22] B. Kingsbury, “Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments,” PhD thesis in Computer Science, University of California, Berkeley, CA, 1998.

[23] P. Hatziantoniou, I. Potamitis, N.-A. Tatlas et al., “Robust speech recognition in reverberant environments based on complex-smoothed responses,” in Speech and Computer International Workshop, Patras, Greece, 2004, pp. 107-110.

[24] H.-M. Park, and R. Stern, "Missing Feature Speech Recognition Using Dereverberation," IEEE, 2007.

[25] A. Shamsoddini, and P. N. Denbigh, “A sound segregation algorithm for reverberant conditions,” Speech Communication, vol. 33, pp. 179-196, 2001.

[26] B. Gillespie, and L. Atlas, “Strategies for Improving Audible Quality and Speech Recognition Accuracy of Reverberant Speech,” in ICASSP, 2003, pp. 676-679.

[27] D. Gelbart, and N. Morgan, “Evaluating Long-term Spectral Subtraction for Reverberant ASR,” in IEEE Workshop on Automatic Speech Recognition and Understanding, 2001, pp. 103-106.

[28] D. Gelbart, and N. Morgan, “Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition,” in ICSLP-2002, Denver, CO, USA, 2002.

[29] T. Takiguchi, and M. Nishimura, "Acoustic model adaptation using first order prediction for reverberant speech." pp. 869-872.

[30] C. K. Raut, T. Nishimoto, and S. Sagayama, “Model Adaptation by State Splitting of HMM for Long Reverberation,” in INTERSPEECH-2005, 2005, pp. 277-280.

[31] N. Roman, S. Srinivasan, and D. Wang, “Binaural segregation in multisource reverberant environments,” J. Acoustical Society of America, vol. 120, no. 6, pp. 12, 2006.

[32] B. Milner, “Comparison of Front-end Configurations for Robust Speech Recognition,” in International Conference on Acoustics Speech and Signal Processing, 2002, pp. 797-800.

[33] R. Stern, F.-h. Liu, Y. Ohshima et al., “Multiple Approaches to Robust Speech Recognition,” 2006.

[34] L. Deng, J. Wu, J. Droppo et al., "Analysis and Comparison of Two Speech Feature Extraction/Compensation Algorithms," 6, 2005, pp. 477-480.

[35] V. Ahkuputra, S. Jitapunkul, E. Maneenoi et al., “Comparison of Different Techniques on Thai Speech Recognition,” in IEEE Asia-Pacific Conference on Circuits and Systems, 1998.

[36] C. Guan, C. Zhu, Y. Chen et al., “Performance Comparison of Several Speech Recognition Methods,” in International Symposium on Speech, Image Processing and Neural Networks, Hong Kong, 1994, pp. 710-713.

[37] R. Flynn, and E. Jones, “A Comparative Study Of Auditory-Based Front-Ends For Robust Speech Recognition Using The Aurora 2 Database,” in Irish Signal and Systems Conference, Dublin Institute of Technology, 2006.

[38] Y. Pan, and A. Waibel, “The Effects of Room Acoustics on MFCC Speech Parameter,” in ICSLP, Beijing, China, 2000, pp. 129-132.

[39] "The HTK Book," S. Young, ed., Cambridge University Engineering Department, 2001-2006.

[40] "Dragon NaturallySpeaking 9 Preferred," http://www.nuance.com/naturallyspeaking/preferred/.

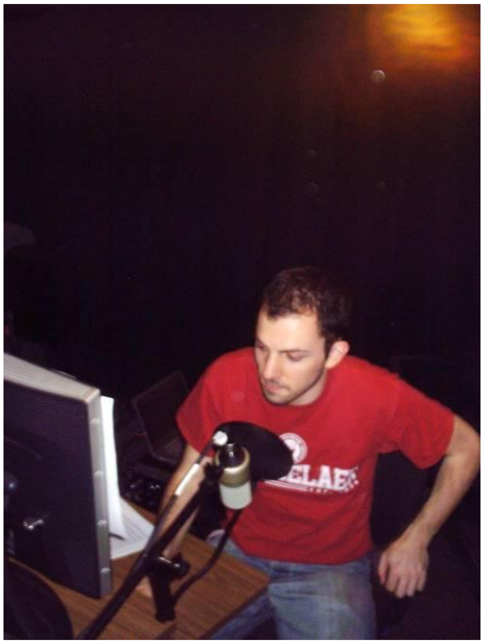
[41] L. Rosa. "Speech Code," http://www.advancedsourcecode.com/.

[42] J. Braasch, "Localization in the Presence of a Distracter and Reverberation in the Frontal Horizontal Plane: II. Model Algorithms," Acta Acustica, vol. 88, pp. 956-969, 2002.

[43] M. Bakke, "LexSTI," Lexington Center and School for the Deaf, 2003.

[44] H. J. M. Steeneken, and T. Houtgast, "A physical method for measuring speech-transmission quality," J. Acoustical Society of America, vol. 67, pp. 318-326, 1980.

A. EXPERIMENT PICTURES



B. MODIFIED RHYME TEST

8.3 Modified Rhyme Test

The Modified Rhyme Test (MRT) consists of 50 six-word lists of monosyllabic English words (House, 1965). Almost all the words have three sounds in a consonant-vowel-consonant sequence. The listeners are offered six words from which to choose the correct one. A carrier sentence may be used in which the test word is spoken without emphasis. The correct response is always offered. The alternatives differ only in either the initial or the final consonant sound. A visual presentation of the listeners' alternative responses, including the stimulus word, shall always be provided to the listener prior to the auditory presentation of the stimulus word.[3] Examples in the MRT are to choose among sun, nun, gun, fun, bun, or run or among peace, peas, peak, peal, peat, or peach. The lists of words are given in Table 2.

TABLE 2. The 300 Stimulus Words Used in the Modified Rhyme Test (MRT) Arranged According to Response Ensembles. (Each of the six words in a response ensemble can serve as the stimulus word for that ensemble. The 50 response ensembles can be randomized to provide different test lists, and various word arrangements within ensembles can be used to prevent possible spatial biases in response.) (From House, 1965).

1	went	sent	bent	dent	tent	rent
2	hold	cold	told	fold	sold	gold
3	pat	pad	pan	path	pack	pass
4	lane	lay	late	lake	lace	lame
5	kit	bit	fit	hit	wit	sit
6	must	bust	gust	rust	dust	just
7	teak	team	teal	teach	tear	tease
8	din	dill	dim	dig	dip	did
9	bed	led	fed	red	wed	shed
10	pin	sin	tin	fin	din	win
11	dug	dung	duck	dud	dub	dun
12	sum	sun	sung	sup	sub	sud
13	seep	seen	seethe	seek	seem	seed
14	not	tot	got	pot	hot	lot
15	vest	test	rest	best	west	nest
16	pig	pill	pin	pip	pit	pick
17	back	bath	bad	bass	bat	ban
18	way	may	say	pay	day	gay
19	pig	big	dig	wig	rig	fig
20	pale	pace	page	pane	pay	pave
21	cane	case	cape	cake	came	cave
22	shop	mop	cop	top	hop	pop
23	coil	oil	soil	toil	boil	foil
24	tan	tang	tap	tack	tam	tab
25	fit	fib	fizz	fill	fig	fin
26	same	name	game	tame	came	fame
27	peel	reel	feel	eel	keel	heel
28	hark	dark	mark	bark	park	lark
29	heave	hear	heat	heal	heap	heath
30	cup	cut	cud	cuff	cuss	cub
31	thaw	law	raw	paw	jaw	saw
32	pen	hen	men	then	den	ten
33	puff	puck	pub	pus	pup	pun
34	bean	beach	beat	beak	bead	beam
35	heat	neat	feat	seat	meat	beat
36	dip	sip	hip	tip	lip	rip
37	kill	kin	kit	kick	king	kid
38	hang	sang	bang	rang	fang	gang
39	took	cook	look	hook	shook	book
40	mass	math	map	mat	man	mad
41	ray	raze	rate	rave	rake	race
42	save	same	sale	sane	sake	safe
43	fill	kill	will	hill	till	bill
44	sill	sick	sip	sing	sit	sin
45	bale	gale	sale	tale	pale	male
46	wick	sick	kick	lick	pick	tick
47	peace	peas	peak	peach	peat	peal
48	bun	bus	but	bug	buck	buff
49	sag	sat	sass	sack	sad	sap
50	fun	sun	bun	gun	run	nun

C. TABLE RESULTS FOR ACCURACY VS. ABSORPTION MATLAB

Set	Word						Accuracy given Abosrption Percent																					
#	1	2	3	4	5	6	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	%
Different Starting Consonant:																												
1	went	sent	bent	dent	tent	rent	0.17	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
2	hold	cold	told	fold	sold	gold	0.17	0.5	0.83	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
5	kit	bit	fit	hit	wit	sit	0.17	0.33	0.83	0.67	0.83	1	0.83	1	1	1	1	1	0.83	0.83	1	1	1	1	1	1	1	
6	must	bust	gust	rust	dust	just	0.17	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
9	bed	led	fed	red	wed	shed	0.17	0.5	0.67	1	0.83	0.67	1	0.67	0.83	1	1	1	1	1	0.83	1	1	1	1	1	1	
10	pin	sin	tin	fin	din	win	0.17	0.33	0.67	0.83	0.67	0.67	0.83	0.83	0.67	0.83	1	1	1	1	0.83	1	1	1	1	1	1	
14	not	tot	got	pot	hot	lot	0.17	0.33	0.83	0.67	0.67	0.83	0.83	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	
15	vest	test	rest	best	west	nest	0.17	0.5	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
18	way	may	say	pay	day	gay	0.17	0.67	1	1	1	1	1	1	1	1	0.83	0.83	1	1	0.83	1	1	1	1	1	1	
19	pig	big	dig	wig	rig	fig	0.17	0.17	0.83	0.83	0.83	0.67	0.83	1	0.67	1	1	0.83	0.83	0.83	0.83	1	1	1	0.83	1	1	
22	shop	mop	cop	top	hop	pop	0.17	0.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
23	coil	oil	soil	toil	boil	foil	0.17	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
26	same	name	game	tame	came	fame	0.17	0.67	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
27	peel	reel	feel	eel	keel	heel	0.17	0.83	1	1	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
28	hark	dark	mark	bark	park	lark	0.17	0.67	0.67	0.67	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
31	thaw	law	raw	paw	jaw	saw	0.17	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
32	pen	hen	men	then	den	ten	0.17	0.5	0.83	1	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
35	heat	neat	feat	seat	meat	beat	0.17	0.83	1	0.83	0.83	1	1	1	0.83	0.83	1	0.83	1	1	1	1	1	1	1	1	1	
36	dip	sip	hip	tip	lip	rip	0.17	0.33	1	1	0.67	0.67	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
38	hang	sang	bang	rang	fang	gang	0.17	0.83	1	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
39	took	cook	look	hook	shook	book	0.17	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
43	fill	kill	will	hill	till	bill	0.17	0.5	0.83	0.83	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
45	bale	gale	sale	tale	pale	male	0.17	0.67	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
46	wick	sick	kick	lick	pick	tick	0.17	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
50	fun	sun	bun	gun	run	nun	0.17	0.83	1	1	1	1	1	1	1	0.83	1	1	1	1	1	1	1	1	1	1	1	
						Avg:	0.17	0.58	0.89	0.92	0.90	0.94	0.97	0.97	0.96	0.99	0.99	0.98	0.99	0.99	0.97	1.00	1.00	1.00	0.99	1.00	1.00	
						Var:	0.00	0.04	0.02	0.01	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Different Ending Consonant:																												
3	pat	pad	pan	path	pack	pass	0.17	0.5	1	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
4	lane	lay	late	lake	lace	lame	0.17	0.5	0.67	0.67	0.83	0.83	0.83	1	0.83	1	1	0.83	1	0.83	1	0.83	1	1	1	1	1	
7	teak	team	teal	teach	tear	tease	0.17	0.83	1	1	1	1	1	1	1	1	0.83	1	1	1	1	1	1	1	1	1	1	
8	din	dill	dim	dig	dip	did	0.17	0.17	0.17	0.67	0.5	0.33	0.5	0.33	0.5	0.67	0.67	0.67	0.67	1	1	0.83	1	1	1	1	1	
11	dug	dung	duck	dud	dub	done	0.17	0.33	0.83	0.83	0.67	0.83	0.83	0.83	1	0.83	0.83	1	1	1	1	1	1	1	1	1	1	
12	sum	sun	sung	sup	sub	sud	0.17	0.67	1	1	0.83	0.83	1	1	0.83	1	1	1	1	1	1	1	1	1	1	1	1	
13	seep	seen	seethe	seek	seem	seed	0.17	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
16	pig	pill	pin	pip	pit	pick	0.17	0.67	1	1	1	0.83	1	1	1	1	1	1	1	1	0.83	1	1	1	1	1	1	
17	back	bath	bad	bass	bat	ban	0.17	0.83	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
20	pale	pace	page	pane	pay	pave	0.17	0.67	0.83	1	1	0.83	0.83	1	0.83	1	1	1	1	1	1	0.83	1	1	1	1	1	
21	cane	case	cape	cake	came	cave	0.17	0.67	0.83	1	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
24	tan	tang	tap	tack	tam	tab	0.17	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
25	fit	fib	fizz	fill	fig	fin	0.17	0.67	0.83	0.83	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
29	heave	hear	heat	heal	heap	heath	0.17	0.83	0.67	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
30	cup	cut	cud	cuff	cuss	cub	0.17	0.67	0.5	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
33	puff	puck	pub	pus	pup	pun	0.17	0.5	0.67	1	1	1	1	1	1	1	1	1	1	1	1	0.83	1	1	1	1	1	
34	bean	beach	beat	beak	bead	beam	0.17	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
37	kill	kin	kit	kick	king	kid	0.17	0.5	0.67	0.67	0.83	0.67	1	1	0.83	1	1	1	1	1	0.83	1	1	1	1	1	1	
40	mass	math	map	mat	man	mad	0.17	0.83	0.67	0.83	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
41	ray	raze	rate	rave	rake	race	0.17	0.33	0.67	0.83	0.5	0.83	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
42	save	same	sale	sane	sake	safe	0.17	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
44	sill	sick	sip	sing	sit	sin	0.17	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
47	peace	peas	peak	peach	peat	peal	0.17	0.67	0.83	0.83	0.83	0.83	0.83	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	
48	bun	bus	but	bug	buck	buff	0.17	0.5	0.67	0.67	0.67	0.83	0.67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
49	sag	sat	sass	sack	sad	sap	0.17	0.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
						Avg:	0.17	0.63	0.81	0.89	0.89	0.91	0.93	0.96	0.95	0.98	0.97	0.98	0.99	0.99	0.99	0.97	1.00	1.00	1.00	1.00	1.00	
						Var:	0.00	0.03	0.04	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Different Vowel Sound:																												
51	when	wan	wane	woan	won	ween	0.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
52	grin	gran	groan	gren	green	groon	0.2	0.8	1	0.8	1	0.8	1	0.8	1	1	1	1	1	1	1	1	1	1	1	1	1	
53	loot	late	light	leet	lat	lot	0.3	1	1	1	1	0.8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
54	miss	mass	moss	mess	muss	moose	0.2	1	1	0.8	0.8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
55	tap	toap	teep	tip	tup	type	0.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
56	shag	shig	shog	sheeg	shoog	shayg	0.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
						Avg:	0.19	0.97	1.00	0.94	0.97	0.94	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
						Var:	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
						Net Average	0.17	0.65	0.87	0.91	0.90	0.93	0.96	0.97	0.96	0.98	0.99	0.98	0.99	0.99	0.98	0.99	1.00	1.00	1.00	1.00	1.00	

Absorption (%)	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
Avg. Acc. (%)	17	28	53	63	75	82	88	91	94	93	96	96	97	99	99	99	99	99	99	100	100
Std. Dev. (%)	2	19	22	21	17	16	15	12	11	13	11	9	9	4	6	5	5	3	4	2	0
Diff. start cons. Avg Acc (%)	17	24	47	56	72	83	87	94	97	96	95	97	98	99	99	99	99	100	100	99	100
St. Dev. (%)	0	14	21	18	14	14	15	12	6.8	8.7	10	6.8	5.5	3.3	4.6	4.6	4.6	0	0	3.3	0
Diff. end cons. Avg Acc (%)	17	25	50	62	72	79	86	87	93	89	95	95	95	99	98	98	99	99	97	100	100
St. Dev. (%)	3.3	16	19	21	20	18	16	13	12	16	12	12	12	4.6	7.3	5.5	4.6	3.3	6.2	0	0
Diff. vowel Avg Acc (%)	17	56	83	92	92	94	94	97	92	97	97	100	100	100	100	100	97	97	100	100	100
St. Dev. (%)	0	25	15	9.1	9.1	8.6	8.6	6.8	20	6.8	6.8	0	0	0	0	0	6.8	6.8	0	0	0

D. TABLE RESULTS FOR ACCURACY VS. ABSORPTION MATLAB VARIABLE INPUT

Set	Word						Accuracy given Abosrption Percent											
#	1	2	3	4	5	6	0	10	20	30	40	50	60	70	80	90	100	%
1	went	sent	bent	dent	tent	rent	0.17	0.17	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	
2	hold	cold	told	fold	sold	gold	0.17	0.17	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	
5	kit	bit	fit	hit	wit	sit	0.17	0.17	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	
6	must	bust	gust	rust	dust	just	0.17	0.17	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	
9	bed	led	fed	red	wed	shed	0.17	0.17	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
10	pin	sin	tin	fin	din	win	0.17	0.17	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	
14	not	tot	got	pot	hot	lot	0.17	0.17	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	
15	vest	test	rest	best	west	nest	0.17	0.17	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
18	way	may	say	pay	day	gay	0.17	0.17	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	
19	pig	big	dig	wig	rig	fig	0.17	0.17	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	
22	shop	mop	cop	top	hop	pop	0.17	0.17	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	
23	coil	oil	soil	toil	boil	foil	0.17	0.17	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	
26	same	name	game	tame	came	fame	0.17	0.17	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	
27	peel	reel	feel	eel	keel	heel	0.17	0.17	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	
28	hark	dark	mark	bark	park	lark	0.17	0.17	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
31	thaw	law	raw	paw	jaw	saw	0.17	0.17	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	
32	pen	hen	men	then	den	ten	0.17	0.17	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	
35	heat	neat	feat	seat	meat	beat	0.17	0.17	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	
36	dip	sip	hip	tip	lip	rip	0.17	0.17	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	
38	hang	sang	bang	rang	fang	gang	0.17	0.17	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	
39	took	cook	look	hook	shook	book	0.17	0.17	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	
43	fill	kill	will	hill	till	bill	0.17	0.17	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	
45	bale	gale	sale	tale	pale	male	0.17	0.17	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	
46	wick	sick	kick	lick	pick	tick	0.17	0.17	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	
50	fun	sun	bun	gun	run	nun	0.17	0.17	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
							0.17	0.17	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	
51	when	wan	wane	woan	won	ween	0.17	0.17	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	
52	grin	gran	groan	gren	green	groon	0.17	0.17	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	
53	loot	late	light	leet	lat	lot	0.17	0.17	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	
54	miss	mass	moss	mess	muss	moose	0.17	0.17	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	
55	tap	toap	teep	tip	tup	type	0.17	0.17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
56	shag	shig	shog	sheeg	shoog	shayg	0.17	0.17	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	
							0.17	0.17	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	
3	pat	pad	pan	path	pack	pass	0.17	0.17	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	
4	lane	lay	late	lake	lace	lame	0.17	0.17	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
7	teak	team	teal	teach	tear	tease	0.17	0.17	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	
8	din	dill	dim	dig	dip	did	0.17	0.17	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	
11	dug	dung	duck	dud	dub	done	0.17	0.17	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	
12	sum	sun	sung	sup	sub	sud	0.17	0.17	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
13	seep	seen	seethe	seek	seem	seed	0.17	0.17	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	
16	pig	pill	pin	pip	pit	pick	0.17	0.17	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	
17	back	bath	bad	bass	bat	ban	0.17	0.17	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
20	pale	pace	page	pane	pay	pave	0.17	0.17	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	
21	cane	case	cape	cake	came	cave	0.17	0.17	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	
24	tan	tang	tap	tack	tam	tab	0.17	0.17	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	
25	fit	fib	fizz	fill	fig	fin	0.17	0.17	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	
29	heave	hear	heat	heal	heap	heath	0.17	0.17	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	
30	cup	cut	cud	cuff	cuss	cub	0.17	0.17	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
33	puff	puck	pub	pus	pup	pun	0.17	0.17	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	
34	bean	beach	beat	beak	bead	beam	0.17	0.17	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	
37	kill	kin	kit	kick	king	kid	0.17	0.17	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	
40	mass	math	map	mat	man	mad	0.17	0.17	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
41	ray	raze	rate	rave	rake	race	0.17	0.17	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
42	save	same	sale	sane	sake	safe	0.17	0.17	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	
44	sill	sick	sip	sing	sit	sin	0.17	0.17	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	
47	peace	peas	peak	peach	peat	peal	0.17	0.17	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
48	bun	bus	but	bug	buck	buff	0.17	0.17	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
49	sag	sat	sass	sack	sad	sap	0.17	0.17	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
							0.17	0.17	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	
							0.17	0.17	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	
							0.00	0.00	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	

Absorption (%)	0	10	20	30	40	50	60	70	80	90	100
Avg. Acc. (%)	33	57	51	53	55	52	52	53	51	50	51
Variance (%)	13	17	21	21	21	22	21	20	19	18	19

E. TABLE RESULTS FOR ACCURACY VS. ABSORPTION DRAGON

Set	Word						Accuracy given Abosrption Percent																					Anechoic	Original
#	1	2	3	4	5	6	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	100	100
1	went	sent	bent	dent	tent	rent	0.2	0.5	0.3	0.3	0.7	0.5	0.7	1	0.7	0.8	0.5	0.8	1	1	1	1	1	1	1	1	0.8	1	1
2	hold	cold	told	fold	sold	gold	0.2	0	0.2	0	0.2	0.3	0	0	0.5	0.3	0.2	0.3	0.2	0.5	0.7	0.5	0.5	0.3	0.7	0.7	0.3	0.666667	1
3	pat	pad	pan	path	pack	pass	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0.3	0.7	0.5	0.7	0.3	0.8	0.8	0.2	0.666667	0.5
4	lane	lay	late	lake	lace	lame	0	0	0	0.2	0	0.3	0.3	0.2	0.3	0	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.666667	0.833333
5	kit	bit	fit	hit	wit	sit	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.5	0.5	0.3	0	0.5	0.833333
6	must	bust	gust	rust	dust	just	0	0	0	0.2	0.2	0.2	0.3	0.3	0.5	0.7	0.5	0.7	0.5	0.5	0.5	0.7	1	0.8	1	1	0.5	1	0.833333
7	teak	team	teal	teach	tear	tease	0	0	0.2	0.2	0.2	0.3	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.3	0.5	0.666667
8	din	dill	dim	dig	dip	did	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.2	0.2	0.333333	0.666667
9	bed	led	fed	red	wed	shed	0.3	0.3	0	0.3	0.2	0.3	0.5	0.5	0.5	0.5	0.7	0.5	0.8	0.8	1	1	1	1	1	0.8	0.7	1	1
10	pin	sin	tin	fin	din	win	0	0	0	0.2	0.2	0.2	0.2	0.2	0.3	0.5	0.2	0.2	0.3	0	0.5	0.5	0.5	0.3	0.5	0.7	0.3	0.333333	0.5
11	dug	dung	duck	dud	dub	done	0	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.5	0.666667
12	sum	sun	sung	sup	sub	sud	0	0	0	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.5	0.5	0.5	0.5	0.5
13	seep	seen	seethe	seek	seem	seed	0.2	0.2	0	0	0.3	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.3	0.5	0.333333
14	not	tot	got	pot	hot	lot	0	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.5	0.666667
15	vest	test	rest	best	west	nest	0	0	0	0	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.3	0.5	0.7	0.7	0.7	0.7	0.8	0.5	0.833333	1
16	pig	pill	pin	pip	pit	pick	0	0	0	0	0	0	0	0	0.2	0.3	0.2	0.2	0.5	0.2	0.3	0.5	0.5	0.5	0.5	0.3	0.5	0.666667	0.833333
17	back	bath	bad	bass	bat	ban	0.2	0	0	0.2	0.2	0.2	0.2	0.2	0.2	0.5	0	0.3	0.5	0.5	0.7	0.7	0.8	0.8	0.8	0.8	0.3	0.666667	1
18	way	may	say	pay	day	gay	0.3	0.3	0.2	0.3	0.5	0.5	0.5	0.7	0.7	0.7	0.5	0.5	0.5	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.8	1	1
19	pig	big	dig	wig	rig	fig	0	0	0	0	0	0	0	0.2	0	0	0	0.2	0.2	0.3	0.2	0.2	0	0	0	0	0.2	0.166667	0.333333
20	pale	pace	page	pane	pay	pave	0.2	0	0	0.3	0	0.3	0.3	0.2	0.3	0.8	0.3	0.3	0.5	0.3	0.5	0.7	0.7	0.7	0.7	0.8	0.3	1	1
21	cane	case	cape	cake	came	cave	0	0	0	0	0	0.2	0	0.3	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.2	0.3	0.3	0.5	0.2	0.3	0.5	0.666667
22	shop	mop	cop	top	hop	pop	0	0.2	0	0	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.333333	0.5
23	coil	oil	soil	toil	boil	foil	0	0	0	0	0	0.2	0	0.2	0.2	0.2	0.3	0.2	0.5	0.7	0.7	0.8	1	1	1	1	1	1	1
24	tan	tang	tap	tack	tam	tab	0	0	0	0.2	0	0.2	0	0.3	0.3	0.3	0.3	0.2	0.5	0.5	0.7	0.7	0.7	0.7	0.5	0.7	0.5	0.666667	0.833333
25	fit	fib	fizz	fill	fig	fin	0	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.333333	0.5
26	same	name	game	tame	came	fame	0.2	0.2	0.2	0.3	0.3	0.3	0.7	0.5	0.8	0.7	0.5	0.5	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.8	0.833333	0.666667
27	peel	reel	feel	eel	keel	heel	0.2	0.2	0	0.2	0.5	0.3	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.833333	0.833333
28	hark	dark	mark	bark	park	lark	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.5	0.3	0.5	0.5	0.5	0.5	0	0.666667	0.666667
29	heave	hear	heat	heal	heap	heath	0.3	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.2	0.3	0.3	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.333333	0.833333
30	cup	cut	cud	cuff	cuss	cub	0	0	0	0	0	0	0	0	0.2	0	0	0.2	0.2	0	0.3	0.2	0.3	0.5	0.3	0.5	0	0.666667	0.833333
31	thaw	law	raw	paw	jaw	saw	0.2	0	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0.333333	0.333333
32	pen	hen	men	then	den	ten	0.2	0	0.2	0	0.2	0.3	0.3	0.3	0.3	0.3	0.5	0.3	0.3	0.3	0.5	0.5	0.3	0.5	0.5	0.5	0.3	0.5	0.5
33	puff	puck	pub	pus	pup	pun	0	0	0	0	0	0	0	0	0.2	0.2	0.2	0	0	0.2	0.2	0.2	0.3	0.3	0.3	0.2	0.2	0.333333	0.5
34	bean	beach	beat	beak	bead	beam	0	0	0	0	0	0	0	0	0	0.2	0	0.2	0	0.2	0.2	0.2	0.2	0.2	0	0.2	0.2	0.166667	0.333333
35	heat	neat	feat	seat	meat	beat	0.2	0	0.2	0	0.2	0.2	0.3	0	0.2	0.3	0.2	0.2	0.5	0.3	0.5	0.7	0.5	0.7	0.5	0.7	0.8	0.5	0.833333
36	dip	sip	hip	tip	lip	rip	0	0	0	0	0	0	0	0.2	0.3	0	0.2	0.3	0.2	0.3	0.3	0.3	0.5	0.3	0.3	0.3	0.2	0.833333	0
37	kill	kin	kit	kick	king	kid	0.2	0	0	0	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.7	0.5	0.5	0.7	0.8	0.8	0.7	0.833333	0.833333
38	hang	sang	bang	rang	fang	gang	0	0	0	0.2	0.3	0.5	0.3	0.5	0.5	0.5	0.5	0.7	0.5	0.7	0.7	0.5	0.5	0.7	0.8	0.7	1	0.666667	1
39	took	cook	look	hook	shook	book	0	0	0.2	0	0	0	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.5	0.5	0.5	0.7	0.7	0.2	0.666667	0.666667
40	mass	math	map	mat	man	mad	0	0	0	0.3	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.166667	1
41	ray	raze	rate	rave	rake	race	0	0.2	0.2	0.3	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.5	0.5	0.3	0.5	0.7	0.7	0.7	0.2	0.666667	0.666667
42	save	same	sale	sane	sake	safe	0.2	0	0.2	0.3	0.5	0.3	0.3	0.2	0.2	0.2	0.5	0.2	0.3	0.3	0.2	0.5	0.7	0.3	0.5	0.5	0.5	0.5	0.666667
43	fill	kill	will	hill	till	bill	0	0	0	0.2	0.3	0.3	0.5	0.3	0.7	0.5	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.666667	0.666667
44	sill	sick	sip	sing	sit	sin	0	0	0	0	0	0.2	0.2	0.2	0.3	0.3	0.2	0.2	0.2	0.2	0.3	0.2	0.5	0.3	0	0.5	0.2	0.666667	0.833333
45	bale	gale	sale	tale	pale	male	0	0	0	0.2	0.2	0.7	0.5	0.7	0.8	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.833333	0.833333
46	wick	sick	kick	lick	pick	tick	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.2	0	0	0.2	0.2	0.2	0.2	0.666667	0.666667
47	peace	peas	peak	peach	peat	peal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.166667	0.333333
48	bun	bus	but	bug	buck	buff	0	0	0	0	0	0	0	0	0	0.3	0	0.2	0.3	0.3	0.3	0.5	0.7	0.3	0.8	0.7	0	1	1
49	sag	sat	sass	sack	sad	sap	0	0.2	0	0.2	0	0.2	0.2	0.2	0.3	0.2	0.2	0.3	0.3	0.5	0.3	0.2	0.3	0.3	0.3	0.5	0.3	0.666667	1
50	fun	sun	bun	gun	run	nun	0.2	0	0	0.3	0.5	0.3	0.7	0.5	0.8	0.8	0.7	0.8	0.8	1	1	1	1	1	1	1	0.8	0.833333	1
51	when	wan	wane	woan	won	ween	0	0	0	0	0	0	0.2	0.2	0.2	0.5	0.3	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.166667	0.5
52	grin	gran	groan	gren	green	groon	0	0	0.2	0.2	0.5	0.2	0.2	0.2	0.3	0.2	0.3	0.3	0.3	0.2	0.2	0.3	0.5	0.3	0.5	0.2	0.3	0.333333	0.5
53	loot	late	light	leet	lat	lot	0	0	0	0	0	0	0	0	0.2	0.2	0.2	0	0	0	0.2	0.2	0.2	0.2	0.2	0	0	0.166667	0.166667
54	miss	mass	moss	mess	muss	moose	0	0	0	0	0	0	0	0.2	0	0	0	0	0.2	0.2	0	0	0	0.2	0.3	0.2	0	0	0.333333
55	tap	toap	teep	tip	tup	type	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.2	0.3	0.2	0.3	0.3	0.3	0.3	0.2	0.5	0.833333
56	shag	shig	shog	sheeg	shoog	shayg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.333333

Absorp. (%)	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	Ane-choic	Orig-inal
Avg. Acc. (%)	5	4	3	10	12	17	19	20	27	27	25	26	31	33	38	39	43	43	46	46	35	55	67
Variance (%)	1	1	0	1	2	2	4	3	5	5	4	4	5	6	6	7	8	7	8	8	8	8	7
Diff. start cons. Avg Acc (%)	8	8	6	10	19	24	27	30	37	35	33	35	40	44	50	53	53	55	59	59	47	68	72
Variance (%)	1	2	1	2	3	3	6	7	9	8	6	7	8	8	8	9	10	9	9	9	11	6	7
Diff. end cons. Avg Acc (%)	5	3	3	10	8	15	14	15	21	23	19	22	25	27	33	32	40	37	42	43	28	55	71
Variance (%)	1	0	0	2	2	1	2	1	2	3	2	1	3	3	4	4	5	4	7	6	3	5	5
Diff. vowel Avg Acc (%)	0	0	3	3	8	3	6	8	11	14	14	14	19	17	19	19	25	25	31	19	17	19	44
Variance (%)	0	0	0	0	4	0	1	1	2	4	3	5	4	3	4	4	5	3	4	4	4	4	5