Testing the Limits – Quantifying the Degradation of Automatic Speech Recognition in Reverberant Environments

Stephen Secules [1], Jonas Braasch[2]

[1] Arup Acoustics, London, United Kingdom
[2] Faculty of Architecture, Rennselaer Polytechnic Institute, Troy, New York, United States
Stephen.secules@arup.com, braasj@rpi.edu

Abstract

Surprisingly little is known about the specific character of the depreciation of automatic speech recognition (ASR) in reverberation— its primary acoustics causes (e.g. room geometry, reverberation strength) or speech 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 recognition task of similar sounding word lists. A range of reverberant settings was simulated by convolution with an impulse response. 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.

Index Terms: automatic speech recognition, reverberation, comparative analysis

1. Introduction

The main barrier to an ASR implementation in new applications is source quality. An ASR system is mainly 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.

While ASR was born out of speech and acoustics sciences research, the two fields often operate on separate but parallel planes. With regards speech in reverberation, the acoustics sciences community is concerned with arriving at a more precise definition of the problem. The ASR community takes a largely engineering-based iterative design approach which is leading towards a maximized efficiency of the current methods. The two fields have much more they could share: a well-documented and scientific characterization of the specific problems facing ASR would 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 of further collaboration 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.

As part of this overall goal, the focus of 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 adds 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.

This paper begins with an introduction to the motivation and research focus of the study. It continues with a summary of the room acoustics research into human speech perception, as well as a review of the treatment of room acoustics in ASR research. 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 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 [1]. 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.

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 constant fundamental frequency and its harmonics are removed by a process called *cepstral smoothing*. 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. 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. 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. The resultant frequency spectrum is a representation of only short term spectral attributes of speech [1]. This spectral magnitude forms the mel-frequency cepstral coefficient (MFCC) at each frequency, in the same way that Fourier coefficients represent linear spectral magnitude.

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 [2]. 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.

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) [3]. 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. Rather than a direct Markov process, 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.

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.

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. 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 [4].

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 implemen-tation of speech recognition today.

3. Treatment of Room Acoustics in Current Research

One of the problems with addressing the area of ASR in reverberant settings is the lack of precise room acoustics science in the discussion. Although many studies document an improvement in reverberant speech recognition based on an implementation strategy, they may not include 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 a selection of ASR 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 nature of the reverberation added. Takiguchi et al. only cite the source of the impulse response used and list its length, not giving any other relevant information on its properties [5]. The room acoustics-focused study by Pan et al., 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" [6]. Shamsoddini presents a two microphone segregation technique which uses temporal cues to imitate the precedence effect and harmonicity cues of the voice signal to suppress non-harmonic noise. The methodology reported only 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 [7]. 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 studies 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, listing the reverberation time of one of them [8]. Gillespie's research on dereverberation shows the accuracy results for 6 reverberation times as they have removed some of the uncorrelated non-speech energy [9]. The Park et al. binaural dereverberation research analyzes the effect of two different reverberation times and two signal-to-interferer ratios [10]. Hatziantoniou tests the HTK toolbox recognizer with two real room settings measured directly with impulse response techniques in a classroom and reports the reverberation time and volume of the rooms [11]. Roman's tests use linearly increased reverberation times to analyze the performance of their binaural segregation algorithm on suppressing noise in reverberant settings, though their methodology does not explore the effect of the reverberation times in the main ASR experiment [12]. Finally, Palomäki et al. have included a thorough evaluation of their methodology over a wide variety of specified room acoustics environments [13]. 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 would still make the results hard to compare.

As shown above, there is an issue of breadth, precision, and consistency 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. 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.

4. 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. Automatic Speech Recognition Platform

Two ASR platforms were tested in this research: Dragon Naturally Speaking® (commercially available recognizer) and a Matlab toolbox developed by Luigi Rosa [14]. The toolbox 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 may 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. As a first step into exploring the exact nature of the effect of reverberation on speech recognition platforms; the basic experimental method outlined in Section 4.1 could easily be applied to other platforms in the future.

4.2. 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. 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.

4.3. Black Box Investigation Procedure

The room acoustic settings for this experiment are modeled using a geometrical acoustics platform implemented by Braasch [15]. As previously described, 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. Although arbitrarily chosen, these parameters were thought to be fairly representative of the dimensions of a midsized conference room, a fairly typical speech recognition environment.

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 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.

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

Table 1: Absorption Coefficients and Room Acoustic Parameters

5. Results

Below are the results using the Matlab platform to investigate the overall speech recognition accuracy in reverberation. The dotted line at 17% accuracy reflects the performance of random guessing. The average accuracy percentage for each absorption and corresponding reverberation time is graphed here, with the error bars representing a standard deviation measure above and below. Also shown in this figure are the results of the comparison of impulse response components: early reflections and reverberant tail.

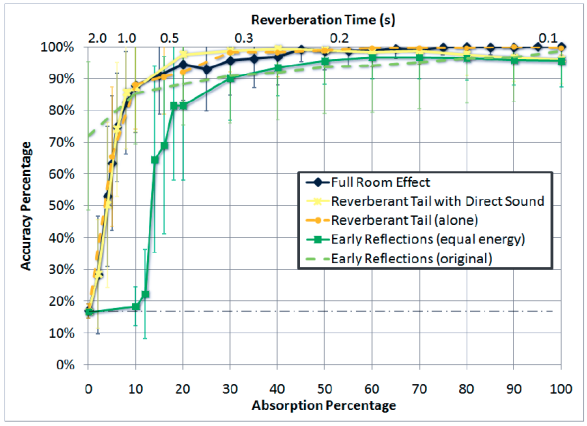

Figure 1: Impulse Response Components

The shape of the response from the full room effect is similar to the response from both reverberant tail and reverberant tail with direct sound. These curves all show a steep drop off past 1s reverberation time, which is similar to the effective drop off point for human speech intelligibility. The response from the strength of early reflections is different from the other curves. While the original early reflections vector shows an increased performance at 2 second reverberation, when normalized to have equal strength as the reverberant tail it shows a dramatically decreased performance.

Figure 2 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.

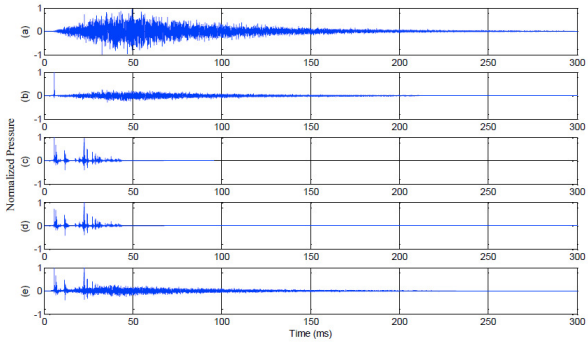

Figure 2: Impulse Response Comparison Absorption 20%

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.

The overall results of the study were compared with five other ASR studies which reported reverberation properties or reverberation times in their results. Figure 3 shows the Matlab platform and Dragon NaturallySpeaking results, compared with each of the Palomäki, Hatziantoniou, Park, Gillespie, and Gelbart data.

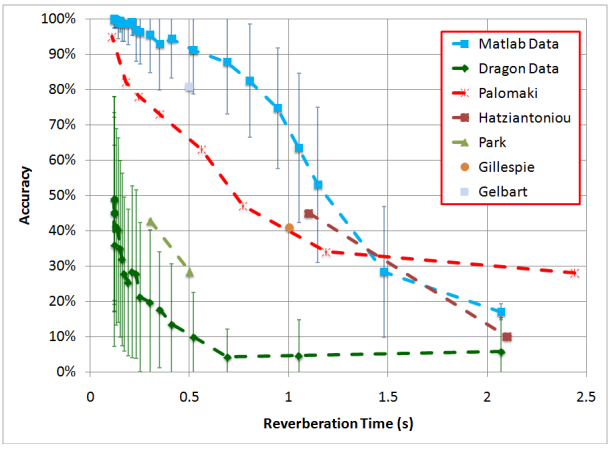

Figure 3: Comparison of Reverberant ASR Studies

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 than most data gathered from other studies, while the Dragon recognizer is less robust. 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.

6. Conclusions

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. 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.

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.

7. Future Research

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 com-munity 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.

8. Acknowledgements

The author wishes to express thanks to Arup for funding his attendance here.

9. References

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

[2] 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.

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

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

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

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

[7] A. Shamsoddini, and P. N. Denbigh, "A sound segregation algorithm for reverbe-rant conditions," Speech Communication, vol. 33, pp. 179-196, 2001.

[8] 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.

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

[10] H.-M. Park, and R. Stern, "Missing Feature Speech Recognition Using Derever-beration," IEEE, 2007.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.