Final Report: OUCH Project

(Outing Unfortunate Characteristics of HMMs)

Nelson Morgan*, Jordan Cohen§, Sree Hari Krishnan Parthasarathi*, Shuo-Yiin Chang*, and Steven Wegmann*

TR-13-006

September 2013

This is a final report submitted for the research effort supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory contract number FA8650-12-C-7217. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

* International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, California, 94704
§ Spelamode Consulting

TABLE OF CONTENTS

Section

LIST OF FIGURES

LIST OF TABLES

1.0 EXECUTIVE SUMMARY – WHAT'S WRONG WITH ASR, AND HOW CAN WE FIX IT?

Automatic speech recognition (ASR) forms a critical link in the acquisition of information from audio and video data. Currently, the accuracy of this component in common real world acoustic conditions is quite poor. Depending on acoustic conditions and microphone placement, speech recognition error rates for conversational speech range from the mid-teens to 30-50%, even for the best systems. This range makes further analysis by humans or machines extremely difficult.

Over the last year, with sponsorship from IARPA and AFRL, we have focused on determining the primary sources of these difficulties. We did this through two separate mechanisms: an in-depth study of the source of errors in the acoustic model, using a novel sampling process to quantify the effects that the two major HMM assumptions have on recognition accuracy; and a broader study of problems in this area, in which we relied on a survey of area experts and of the relevant literature.

In the in-depth study, we have obtained results that demonstrate, among other things, that a lack of robustness (to mismatched training/test conditions) is a significant source of error in our own experiments, and that the sensitivity to such mismatch in the acoustic representations is a prominent source of errors. However, our results also show that in the case of matched conditions, the incorrect assumptions inherent to our standard statistical models is the dominant source of errors.

In particular, by exploiting a resampling method based on Efron's bootstrap [1], we constructed a series of pseudo datasets from near-field and far-field meeting room datasets, that at one end satisfied the HMM model assumptions, while at the other end deviated from the model in the way real data did. Using these datasets we probed the standard HMM/GMM framework for automatic speech recognition. Experiments show that when the conditions are matched (even if they are far-field), the model errors dominate; however, in mismatched conditions features are neither invariant nor are they separable using the near-field models, and contribute as much to the total errors as does the model. We then studied unsupervised MLLR adaptation as a means to compensate for this issue in the model space; while this approach mitigates the errors, the conclusions about the lack of invariance of the MFCC features in mismatched conditions still holds true. As part of future work, this study paves way for principled investigations into other spectro-temporal representations [2].

Our surveys of ASR researchers and of the ASR literature have provided a further sense of the community's perspective on the topic. Our informants believed that they were working with an emerging technology. In fact, there was a note of cynicism from many as they felt that the core recognition models were so old, that the technology had been an emerging technology for 30 years. It was noted as immature in essentially all of the technical aspects of recognition. While there was minor dissatisfaction with recognition performance per se, the major complaint was that the speech recognition systems that are deployed today are not robust to conditions other than the training conditions. They degrade rapidly and not gracefully in noise, for novel speakers, in far-field or other unusual acoustic conditions, in accented speech, and for speech in

which other signals or noises share the acoustic channel.

Our informants identified essentially every element of the current ASR technology as the focus of experiments to attempt to improve the technology. Failures were abundant, and performance continues to lag that of people in similar situations.

Our two studies were primarily focused on finding the source of difficulties in ASR technology, and the determination of promising directions is much harder. That being said, given the extremely low error rates for data matching our models' independence assumptions, it is likely that explorations of methods for properly representing the conditional dependence between frames and phones (given the state) should have a major effect. On the other hand, given the problems that our community identified with brittle systems and their lack of robustness, our results point to the relevance of acoustic representations that would be more invariant to such mismatches, or those that easily compensate for those conditions. Furthermore, the use of resampling techniques such as the ones we have used could provide a useful tool during the development of methods to handle these two issues – it could provide a more sensitive indicator than just looking at the word error rate for the real data.

2.0 INTRODUCTION

2.1 Historical Background

Speech recognition is defined as the science of recovering words from an acoustic signal meant to convey those words to a human listener. Since the initial use of patterns in speech displayed by "spectrograms", developed during World War II but released to the public in the years following the war, the art of speech recognition has gone through several phases. Early work centered on hand-crafted models of spectra and their movements, such as the early digit recognizers from Bell Laboratories. In the 1970's recognition work was focused on Dynamic Time Warping (DTW), where some spectral distance was coupled with a time-warping algorithm and the space of potential warps was searched using a dynamic program. In the 1970's, the Hidden Markov Model (HMM) approach was developed by Jim Baker at Carnegie Mellon, and by Fred Jelinek and his team at IBM, following fundamental developments by a small number of research scientists at the Institute for Defense Analysis (IDA) in Princeton, NJ.

The IDA team brought the community together in 1982, in a seminar in Princeton, NJ, where they outlined the benefits and practice of HMMs. Shortly thereafter, the Bell Laboratories team, under Larry Rabiner, published several papers comparing the results of speech recognition using DTW and HMM models, noting the substantial improvements of HMM over DTW systems. The field then pivoted to HMM systems. Despite the earlier developments at other laboratories noted above, it was the Bell Laboratories publications that swayed the community at large to use HMMs.

Additionally, DARPA funded several projects in speech recognition, from the earliest in the 1970's to the latest in the 2000's. While the early projects focused on technology, later projects emphasized pushing the existing technology into more challenging areas, and creating systems

that worked in noisy, distorted, and spontaneous conditions. In addition, there was an attempt to create speech-to-speech translation in the Global Autonomous Language Exploitation (GALE) project, where the recognizer simply created a word string that was then manipulated to form words in the target language. Arguably, none of the later projects focused on improving the underlying technology of speech recognition.

Commercialization of the technology has been successful in Interactive Voice Response (IVR) systems with limited vocabularies that provide self-service options for customers calling into contact centers, and in dictation products with motivated, engaged talkers. However, more complex, natural language IVR applications have required costly professional services engagements in order to tightly tune the applications to work. Additionally, ASR has not been successful for general transcription applications either, as error rates have remained stubbornly high. The advent of powerful smart phones with high quality audio systems, internet connections, and substantial computing power has created a new interest in speech recognition technology commercially. For example, as of 2012, a multitude of vendors providing contact center and customer service applications have developed speech-enabled customer care applications on smart phones. While these work well in many environments, they often fail in accented speech, in noisy situations, and in other challenging acoustic environments. Overall, NIST, who tracks performance of government funded speech recognition systems, has found that the tremendous decrease in error rates seen in the '70s and '80s has slowed to a crawl, and in fact the primary improvements they have reported in the last decade were with fairly structured data (e.g., Mandarin broadcasts) while reported improvements have slowed to a crawl in less structured tasks such as the transcription of natural meetings.

There has always been a sense among the researchers in speech recognition that the modeling assumptions in HMM systems were too simplistic to be sensible. Larry Gillick and Steven Wegmann, working at Nuance in 2009, explored the hypothesis that the independence assumptions in acoustic models were instrumental in the failure of these models. After this, Steven Wegmann came to ICSI to work on an NSF-funded project to further develop the analysis required, working with Berkeley graduate student Dan Gillick (Larry's son). This early work provided a proof-of-concept to support the current IARPA/AFRL project that examines the issues in some detail, including an analysis of conditions of acoustic mismatch between training and test. Concurrently, the project probes the large commercial and academic community working in speech and language technology to see how they viewed the technology, and if there was an obvious path to an improved technology that was emerging.

Consequently, the two part study for the project is reported on in this document: (1) an in-depth study of the statistical properties of the standard GMM/HMM-based acoustic model, and (2) a breadth-wise study of the overall field based on a community survey and a corresponding literature search.

2.2 In-depth study: The effects of standard HMM assumptions on performance

It is a reasonable hypothesis that one of the major contributing factors to the oft-observed brittleness of ASR is the remarkable inability of the standard HMM-based acoustic model to accurately model speech test data that differs in character from the data that was used for its

training. While there has long been speculation about the root causes of this brittleness, ranging from the over-fitting of the acoustic model to its training data to the lack of invariance of the standard front-end (mel-frequency cepstral coefficients (MFCCs)), there is surprisingly little quantitative evidence available to back up one claim over another. Furthermore, while many authors have explored the circumstances under which recognizers fail (e.g., rapid speech, noise, confusable word pairs, etc.), the research aimed at improving HMM-based speech recognition accuracy has largely ignored questions concerning understanding or quantifying the underlying causes of recognition errors (with some notable exceptions, including [3, 4]). Instead, improvements–many of which are reviewed in [5, 6, 7, 8, 9]–to the front-end and the acoustic models have largely proceeded by trial and error. The research that we will describe benefits from our earlier research described in [10, 11] that used simulation and a novel sampling process to quantify the effects that the two major HMM assumptions have on recognition accuracy. In this previous work, we analyzed recognition performance on tasks[1] where the properties of the training and test acoustic data were not challenging and were homogeneous, or matched, across the training and test sets. In this report, however, we will summarize analysis of recognition performance using the ICSI meeting corpus [12], where the acoustic data are more challenging. In particular, we are able to exploit properties of this corpus to compare recognition performance when the training and test data acoustics are matched or mismatched.

More specifically, we have used the parallel recordings using near- and far-field microphones in the ICSI meeting corpus to construct three sets of related recognition tasks: (a) matched near-field acoustic model training and recognition test data; (b) matched far-field acoustic model training and recognition test data; (c) mismatched near-field acoustic model training data and far-field recognition test data.

We are interested in understanding which properties of real data are surprising to the acoustic models that are in common use in the ASR community, where we will use recognition word error rates as our measure of "surprise". There are many potential sources of surprise (or mismatch) that the data can present to the acoustic models. However, in the in-depth study described here, we are specifically interested in quantifying the effects of the surprise due to statistical dependence in the data, due to the deviation of real data to the parametric form of the marginal distributions in the model (GMMs), and due to training on near-field data and testing on far-field data. In order to obtain accurate estimates of the degree of surprise due to these factors we must, to the extent that it is possible, eliminate other sources of surprise that, while interesting in their own right, are conflating factors in this study. There are two broad categories of factors that we address involving, on the one hand, properties of the data, and, on the other hand, acoustic model technicalities.

2.3 Breadth-wise study: The community/literature surveys on errors in ASR

It was not possible in a one-year study to do a detailed analysis of every potential source of error in automatic speech recognition, certainly not at the level of the in-depth study of the acoustic model introduced above. On the other hand, it was agreed at the outset that we needed to at least consider a broader class of issues in order to better advise the government about fruitful

[1] Based on the Wall Street Journal and Switchboard corpora

directions for future programs. Consequently, we developed a plan to conduct a survey of the speech recognition community and a search of the relevant literature to get a representative sampling of expert opinions on the state of speech recognition. The project included a survey of many of the active participants in the speech recognition community. We wanted to evaluate what people told us about speech technology and its application in the context of our findings about the performance of modern speech recognition algorithms and their flaws. It is our hope that this analysis, in combination with the in-depth study of the statistical acoustic model, will lead to a way forward for the community to improve speech, or at least show us how to analyze the current state of the technology and uncover areas and ideas for future work.

We set out to interview a significant number of major participants in speech and language technology, asking questions about their sense of the technology, their experience with improving the technology, and their projections for the future. Interviews were mostly by telephone, although some were in person, and we followed the "snowball" polling practice (described further in section 3.2), which promised a reasonably unbiased view of international experts' views. We interviewed academics, commercial developers, and government employees. Both through our own intuitions and the suggestions of the interviewees, we also conducted a search of the relevant literature. The results of both of these efforts certainly show a wide diversity of opinions, but there are some major impressions that appear to be justified by the data. Both the common themes and the diversity of opinions are presented in this report, primarily in section 4 (with some description of the participant characteristics in Appendix B).

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

Section 3 describes the methods for both the acoustic model study and the broader survey. Results from these studies will be presented in section 4.

3.1 In-depth study of the acoustic models

3.1.1 Simulation and Resampling Methodology

We used simulation and a novel sampling process to generate pseudo test data that deviate from the major HMM assumptions in a controlled fashion. The novel sampling process, called resampling, was adapted from Bradley Efron's work on the bootstrap [1] in [10, 11]. These processes allowed us to generate pseudo data that, at one extreme, agreed with all of the model's assumptions, and at another extreme, deviated from the model in exactly the way real data do. Across this range, we could control the degree of data/model mismatch. By measuring recognition performance on this pseudo test data, we were able to quantify the effect of this controlled data/model mismatch on recognition accuracy.

3.1.1.1 The simulation and resampling process

The methodology used in this study allows six levels of simulation and resampling: (a) simulation, (b) frame resampling, (c) state resampling, (d) phone resampling, (e) word resampling, and (f) original test utterance.

Simulation: We followed the full generative process assumed by HMMs. The simulated data, therefore, matches all the assumptions of the model. These assumptions are: (a) the sequence of states are hidden and are constrained to follow a Markov chain (b) the features are independent conditioned on the states (c) what the specific form of the probability distribution of the data generated by a given hidden state is. We followed the standard practice in ASR and used Gaussian mixture models (GMMs) for these probability distributions. To generate the test data by simulation, we started with the test transcriptions, and looked up each word in the pronunciation dictionary to create phone transcriptions. We then used the state transitions and the output distribution associated with the states belonging to the triphones to generate the data. Since our feature set has Δ and $\Delta\Delta$ features appended to the static cepstral features and since the GMMs model--and are fit to--the marginal distribution of this feature set, the GMMs never learn about the temporal consistency between a sequence of cepstral vectors and their corresponding $\Delta$and $\Delta\Delta$ features. Consequently, when we simulated from the model the resulting features correspond to the static cepstral features and their Δ and $\Delta\Delta$ features but they have lost (via marginalization) the temporal consistency that the original, real statics, Δ, and $\Delta\Delta$ features had.

Frame resampling: In this case, we did not use the full generative process. Nevertheless, we created data that respects the independence assumptions at different levels. To generate the data in this fashion the following process was performed: (a) We used the training model to perform forced alignment on the training utterances, so that each speech frame is annotated with its most likely generating state. (b) We walked through this alignment, filling an urn for each state with its representative frames; at the end of this process, each urn was populated with frames representing its empirical distribution. (c) To generate resampled data, we used the model to create a forced alignment of the test data, and then sample a frame (at random, with replacement) from the appropriate urn for each frame position; these resampled frames were concatenated. With this frame-level resampling, the pseudo test data was exactly the same length as the original, and had the same underlying alignment, but the frames were then conditionally independent (given the state).

State, phone, and word resampling: By placing entire state sequences of frames in the urns, and then resampling (again, concatenating samples), we ended up with pseudo test data with dependence among frames within state regions, but independence across state boundaries (note that resampling units larger than single frames produces pseudo test data that may be a different length from the original). We further extended this idea to phones and to words; in all cases, the urn labels included the full triphone context. [2]

3.1.1.2 Enforcing common alignment for Near-field and Far-field data

The method of resampling creates an alignment of the training dataset using the recognition model; it then uses the alignments to fill urns that are in turn used to create the pseudo test utterances. The differences in the alignments created by the near-field and the far-field model will lead to the creation of pseudo test sets that are not parallel, leading to the near- field model

[2] Note that while some figures in this document focus on the frame resampling case for simplicity's sake, the tables of results in section 4 and in Appendix A typically show results for all of these levels.

trying to compensate, in addition, for a mismatched alignment. In order to minimize this effect, we created alignments using the near-field model on the near-field data, and used this alignment to generate pseudo, far-field test data (for the mismatched case).

3.1.2 Datasets

We used a dataset of spontaneous meeting speech recorded at ICSI [12] where each spoken utterance was captured using near-field (NF) and far-field[3] (FF) microphones. Our training set was based on the meeting data used for adaptation in the SRI-ICSI meeting recognition system [13]. For the test set we used data from the ICSI meetings drawn from the NIST RT eval sets [14, 15, 16]; this was done to control the variability in the data for the resampling experiments.

The remainder of this subsection discusses the creation of the parallel NF and FF corpora for this project. First, we describe how we estimated and removed a variable length time delay that exists between the corresponding NF and FF utterances, so that each training and test utterance had two parallel versions– NF and FF–that are aligned at the MFCC frame level. This alignment and some further selection were used to choose the specific partitions of parallel NF and FF corpora data to be the training and test sets. These procedures are described in the following subsections.

3.1.2.1 Time-aligning the corpora

In order to synchronize the NF and FF recordings, we must deal with a time delay, or skew, that exists between the two recordings. These time delays arise from two factors: (1) different physical distances between the speakers and the microphones, and (2) systematic delays introduced by the recording software. The latter factor appears to dominate the skew between the NF and FF recordings. Fixed delays were introduced when the channels were initialized at the start of a recording. Since this systematic delay dominates the skew, the NF recordings have a time delay relative to the FF recordings. Fig 1(a) shows audio at the FF microphone that is advanced in time in comparison to the same utterance captured by the NF microphone.

Time delay is more evident in the cross-correlation between the NF and FF signals, as shown in Fig 1(b). The delay could be estimated by searching for a peak in the cross-correlation sequence. In Fig 1(b) the peak is at a lag of 41.88 ms (670 samples at 16 kHz). However, this detection could be difficult because of the recording quality and noise. To guarantee a more precise detection, we divided each utterance into overlapping windows, where the window size was a third of the utterance length and the step size for successive windows was a tenth of the utterance length. For each step, the cross- correlation sequence was calculated and a delay was estimated. If the variation between the estimated delays in the windows for a given utterance was too large, then the estimated delay was regarded as unreliable and the utterance was discarded. Approximately 30% of the utterances were discarded because of these unreliable delay estimates. The delays between NF and FF channels for the reliable data ranged from 12.5 ms to 61.25 ms. This was implemented using the Skewview tool [17]. More about the delay can be found in [18].

[3] We used the SDM or "Single Distant Microphone" recordings for the far-field data.

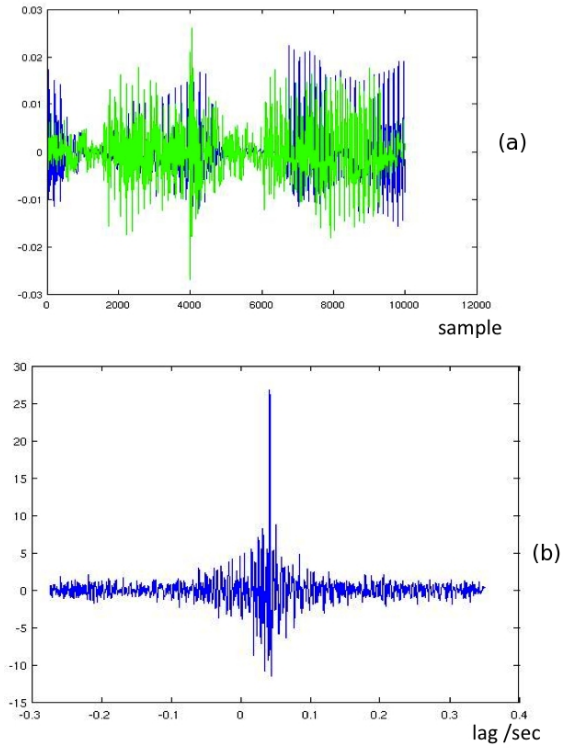

Figure 1. Time alignment – top figure shows near-field in blue, far-field in green, and the bottom figure shows the cross-correlation between the signals

3.1.2.2 Data partitions

Because of the parallel nature of the NF and FF corpora, the data partitions are identical. For simplicity, we describe the NF partitioning. The training set had a dominant speaker accounting for nearly a quarter of the data; this would skew the data generated by the resampling process. On the other hand, perfect speaker balancing cannot be achieved given that this is a corpus of spontaneous speech from natural, unscripted meetings. There is, therefore, a trade-off between amount of data and an egalitarian distribution of speakers. The resulting NF training and test sets consisted of about 20 hours and 1 hour respectively, and their statistics are reported in Table 1.

Table 1. Training and test statistics for near-field (NF) and far-field (FF). The training data is 27.5 hours from the Meeting corpus standard training data. We then removed delay and discarded data that did not survive the deskewing process described above. The test data comprises the ICSI portion of the RT-02, -04s, and -05s evals, after removal of any sentences with OOVs or that did not survive the deskewing process.

Dataset	Speakers	Utterances	Time
Training	26	23729	20.4 (hrs)
Test	18	1063	57.9 (mins)

3.1.3 Models and experimental setup

We used version 3.4 of the HTK toolkit [19] for the front-end, acoustic model training, and decoding. In particular, we used the standard HTK front-end to produce a 39 dimensional feature vector every 10 ms: 13 Mel-cepstral coefficients, including energy, plus their first and second differences. The cepstral coefficients were mean-normalized at the utterance level. We used HDecode for decoding with a wide search beam (300) to avoid search errors. To evaluate recognition accuracy, the reference and the decoded utterances were text normalized and scored using standard NIST tools to obtain word error rates (WERs). The remainder of this section discusses the recognition acoustic models, dictionary, and language model.

3.1.3.1 Near-field acoustic models

The NF acoustic models used cross-word triphones and were estimated using maximum likelihood. Except for silence, each triphone was modeled using a three-state HMM with a discrete linear transition structure that prevents skipping. The output distribution for each HMM state was a GMM with each component being a multivariate Gaussian with diagonal covariance. We used GMMs with 1, 2, 4, and 8 mixture components. While significantly better performance could be achieved with mixtures of more components, the simplicity of a single component is preferable for our analysis; it also highlights the performance differences between our experiments. Maximum likelihood training roughly followed the HTK tutorial: monophone models were estimated from a "flat start", duplicated to form triphone models, clustered to 2500 states and re-estimated.

3.1.3.2 Far-field acoustic models: via single-pass retraining

In one of our key experiments we wanted to isolate and understand how the transformation between the parallel NF and FF data impacts recognition performance when we use NF models to recognize FF data. In order to accomplish this, we wanted to construct parallel NF and FF models whose only differences arise from the transformation between the parallel NF and FF data. Thus, instead of building the FF acoustic models from a flat start, we exploited the parallel nature of the NF and FF training sets to build the FF models using single-pass retraining from the final NF models and the FF data. Single-pass retraining is a form of EM, which is supported by HTK, where, in our case, the E-step is performed using the NF models and data, while the M-step and model updates use the FF data. We only updated the means and variances of the FF models, so the result was a parallel set of NF and FF acoustic models that shared the same state-tying, but the (unknown) transformation between the NF and FF means and variances was determined by the frame-level transformation between the parallel NF and FF acoustic data.

3.1.3.3 Dictionary and language models

Our acoustic models were relatively weak since they were trained from only 20 hours of data, and this was reflected in their small size: up to 8 mixture models per state and only 2500 tied states. However, since the Meeting recognition task is difficult, recognition WERs obtained using the small LM trained from corresponding acoustic training texts are much higher than what is reported in the literature (e.g. in our case 64% in the matched NF condition versus ~30% in the

literature). To ensure more reasonable WERs and more confidence in our results, we used a much more powerful language model. In particular, we used a LM [20] that was trained at SRI by interpolating a number of source LMs; these consisted of webtext and the transcripts of the following corpora: Switchboard, meetings (CMU, ICSI, and NIST), Fisher, Hub4- LM96, and TDT4. We then removed words not in the training dictionary from the trigram LM, and renormalized it. The perplexity of this meeting room LM is around 70 on our test set. In order to use our simulation methodology we need pronunciations for each word in an utterance's transcription. Thus, we removed any test utterances that had any words OOV relative to the SRI dictionary. To be compatible with the SRI LM, we used the SRI pronunciations; that dictionary uses two extra phones in comparison with the CMU phone set –"puh" and "pum"– for hesitations.

3.2 Breadth-wise study

The survey was conducted using "snowball sampling", which is a method for gathering research subjects through the identification of an initial subject or set of subjects who are used to provide the names of other potential subjects [22]. This was used in our study in order to gain access to experts within the field of speech recognition. As such, we started with a few targeted participants, and asked each of them at the end of the survey to give us contact information for two other people within the industry that might participate in our survey.

Whereas the snowball sampling technique can be construed as presenting some bias, in the case of this study we were trying to reach participants with the broadest range of experience within speech recognition. Therefore, having participants nominate those in their peer group they felt had the most experience to draw from was an important factor.

We also asked participants if they would be willing to take a more in-depth survey in the future if we did one.

3.2.1 Demographics

The identities of the interviewees in our survey were anonymized. That is, in keeping with the human subjects requirement from the UC Berkeley IRB, access to the raw subject data was restricted to a limited number of researchers on our team[4]. However, we collected basic demographic information about them to see if we could glean any trend information on who is working in the field.

- Name
- Sex
- Age
- Organization
- Number of Years in Speech Technologies

[4] This was approved by the UC Berkeley Committee for the Protection of Human Subjects, Protocol number 2012-04-4187, April 23, 2012.

- Position/Title
- Questions as to the state of speech recognition

3.2.2 The Questionnaire

The questionnaire was designed to elicit the broadest possible range of answers, and was limited to six questions beyond demographic information.

1. Are you currently working on speech technology products, if so what areas and why?
2. Where has the current technology failed?
3. What do you think is broken?
4. What have you tried to do to improve the technology that should have worked but did not?
5. Why did it fail?
6. Have you solved any speech technology problems that were not published? If so, what?

3.2.3 The Literature Review

As noted in section 2, the community survey was augmented by a review of the relevant research literature. In addition to an abundance of survey articles that were apparent to us, we were also guided by suggestions that came from our interviewees. The key results are given in the next section, and Appendix C provides an additional list of references that either came from our own perspectives or from those of the interviewees.

4.0 RESULTS AND DISCUSSION

4.1 In-depth study of the acoustic models

Near-field (NF) and far-field (FF) test data were created by simulation, and then by resampling frames, states, phonemes, and words; the corresponding recognition models are then used for decoding. Each resampling experiment was jackknifed five times (using different partitions each time for decoding), and the results are shown in the Tables 2, 3, and 4. In the matched NF experiments, NF models were used to recognize NF test data, while the matched FF experiments used FF models and FF test data. In the mismatched experiments, NF models were used to recognize FF test data. Listed in the table for the matched and the mismatched cases were the word error rate (WER), standard error (SE), and the relative increase in WER from previous level of simulation/resampling (the next highest row). The standard errors ranged from 0.03 (simulation in the NF case) to 0.45 (word resampling in the FF case), so all the WER differences between matched and mismatched conditions were significant. Note that the WERs on the test data increased as we move from NF (44.7%) to FF (71.4%), and then to the mismatched conditions (84.7%), that is, for NF models and FF test data; this indicated the difficulty of the tasks.

4.1.1 Analysis of matched near-field results

Near-field results are summarized in Table 2. It is remarkable to see that the WER for simulation and frame resampling is negligibly small in meeting room data, albeit with near-field microphones; for these cases all assumptions made by the model are satisfied by the data. When this is the case, the WER obtained by the system must be similar to human performance. The largest increase in WER is observed when we move from frame resampling to state resampling – a little more than a four-fold increase in errors. Another large increase in WER (123%) occurs when we move down to phone resampling. As dependence is introduced (going down the rows), we start observing larger WER. These results are consistent with what was observed in [11] on the WSJ and Switchboard corpora, both of which also had matched training and test conditions.

Table 2. Rates shown are for simulation from the model and resampling at the different levels of granularity for the near-field matched case. The last column shows the % increase in WER obtained over the next higher level of resampling. All results are for the 1-Gaussian case; similar trends are observed for 8-Gaussian models, but with lower error rates overall (see Appendix A for full results).

Resampling method	WER (%)	Standard Error	Δ WER (%)
Simulation	1.4	0.03	-
Frame	1.9	0.05	31
State	9.6	0.17	416
Phone	21.4	0.21	123
Word	37.6	0.28	75
Original data	44.7	-	19

Figure 2 shows the word error rates for models ranging from 1 to 8 Gaussians per state and for two of the cases shown in the table: resampling at the frame level so that the conditional independence assumptions of the model are satisfied; and the original data, for which these assumptions definitely are not satisfied. Note that the differences in performance due to the number of Gaussians are inconsequential compared to the huge effect of the assumption violation in the original data.

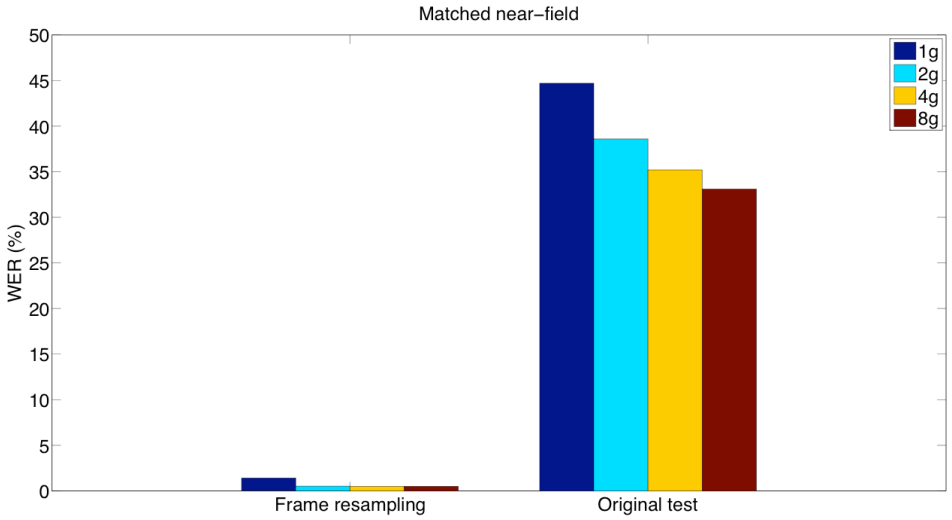

Figure 2. Word error rates for framewise resampled data and for original data, near-field recordings on the ICSI meeting corpus, for triphone models ranging from 1 Gaussian per state to 8 Gaussians per state

4.1.2 Analysis of matched far-field results

Although the WER is consistently worse for the FF results than they were for the NF results, the pattern of error rates over the different resampling methods for the FF case is consistent with what we observe in the NF experiments and in [11]. However, it is striking how small the WER for simulation (1.8%) is when we consider how large the WERs are on real FF data (71.4%). This shows that, when the training and test conditions are matched, and the model assumptions implicit in HMMs are met, MFCC features are essentially separable even for the more challenging FF meeting data.

Table 3. Rates shown are for simulation from the model and resampling at the different levels of granularity for the far-field matched case. The last column shows the % increase in WER obtained over the next higher level of resampling. All results are for the 1-Gaussian case; similar trends are observed for 8-Gaussian models, but with lower error rates overall.

Resampling method	WER (%)	Standard Error	Δ WER (%)
Simulation	1.8	0.03	-
Frame	3.4	0.02	88
State	23.2	0.2	580
Phone	45.5	0.41	96
Word	63.5	0.45	40
Original data	71.4	-	12

Figure 3 shows the word error rates for models ranging from 1 to 8 Gaussians per state and for two of the cases shown in the table: resampling at the frame level so that the conditional independence assumptions of the model are satisfied; and the original data, for which these assumptions definitely are not satisfied. Note that, although the WER for original data is far worse than it was for the near-field data, as with the earlier case, the differences in performance due to the number of Gaussians is inconsequential compared to the huge effect of the assumption violation in the original data.

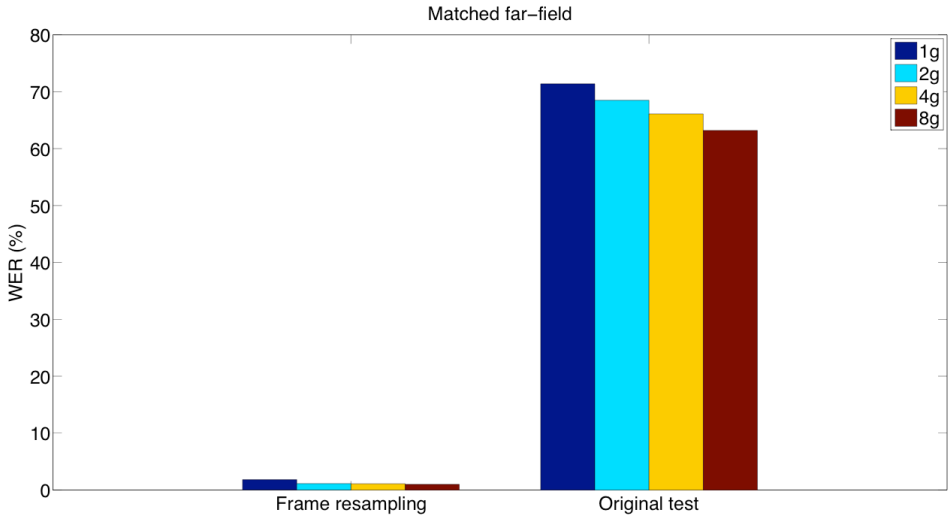

Figure 3. Word error rates for framewise resampled data and for original data, far-field recordings on the ICSI meeting corpus, for triphone models ranging from 1 Gaussian per state to 8 Gaussians per state

4.1.3 Analysis of the mismatched case

The results in the mismatched case are in stark contrast to those obtained for the matched cases. The WER for simulation is much higher at 43%, which indicates that MFCCs are not separable in this mismatched case, i.e., using the near-field models. While the errors due to statistical dependence–the WER from the state resampling to the original data–are considerable (from 59.9% to 84.7%), they are no longer such a dominant cause of recognition errors. To better understand the mismatched simulation result, we compare it to the matched, NF simulation result. In both cases we use NF models to recognize simulated data: in the matched case this data is simulated by the NF models, while in the mismatched case this data is simulated from the FF models. Because we used single-pass retraining to create the FF models from the NF models, the unknown transformation between the NF and FF means and variances is inherited from the unknown transformation between the parallel NF and FF training utterances. Thus the transformation between the test utterances simulated from the NF and FF models is derived from the transformation between the NF and FF models, and it is related to, but much simpler than, the transformation between the parallel NF and FF training data. The NF models have a low WER on the simulated NF test data (1.4%), but they have a high WER (43%) on the simulated FF data,

which is transformed simulated NF data. If the features (MFCCs) were invariant to this transformation, then the WERs would be similar. However, since the WERs are very different, the features cannot be invariant, and the large difference in WERs is due to this lack of invariance.

Table 4. Rates shown are for simulation from the model and resampling at the different levels of granularity for the mismatched case (near-field training, far-field test). The last column shows the % increase in WER obtained over the next higher level of resampling. All results are for the 1-Gaussian case; similar trends are observed for 8-Gaussian models, but with lower error rates overall (see Appendix).

Resampling method	WER (%)	Standard Error	Δ WER (%)
Simulation	43.0	0.23	-
Frame	59.9	0.26	39
State	75.8	0.27	27
Phone	80.6	0.29	6
Word	80.6	0.15	0
Original data	84.7	-	5

Figure 4 shows the word error rates for models ranging from 1 to 8 Gaussians per state and for two of the cases shown in the table: resampling at the frame level so that the conditional independence assumptions of the model are satisfied; and the original data, for which these assumptions definitely are not satisfied. As with the matched cases, the differences in performance due to the number of Gaussians are inconsequential compared to the huge effect of the assumption violation in the original data. However, unlike the matched cases, the error rates for the framewise resampling are not tiny, indicating that even compensating for the conditional dependence in the data does not fix the problem.

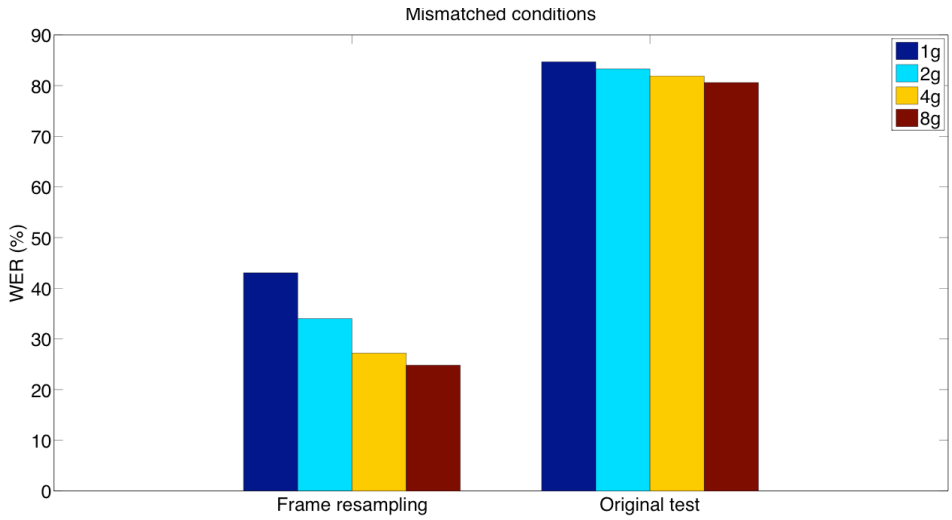

Figure 4. Word error rates for framewise resampled data and for original data, far-field recordings on the ICSI meeting corpus, for triphone models trained on near-field recordings, ranging from 1 Gaussian per state to 8 Gaussians per state

4.1.4 Experiments with some standard methods of improvement

The literature is replete with methods that have been shown to provide incremental reductions in word error rates under various conditions, and it is far beyond the scope of this report to cover all or even a majority of such methods. However, we have experimented with three of the common methods: MLLR adaptation, MPE discriminant retraining, and MLP transformation. Here we describe our results for these three in the context of the experimental methods of this study. The results of the breadthwise studies (given below in section 4.2) will provide a different perspective on the efficacy of the ensemble of such methods. Tables providing specific results for these tests are provided in the Appendix, but the most significant results are described here.

4.1.4.1 Adaptation

A standard approach to mitigating recognition errors due to mismatched conditions is to perform unsupervised MLLR [21], a form of linear mean adaptation. Since the large difference between the matched NF and mismatched simulation results is due to the lack of invariance of MFCCs to a (presumably non-linear) transformation between the NF and FF data, it is natural to try to compensate for this using MLLR. We treat the 1 hour of simulated test data as belonging to a single speaker, and use the recognition hypotheses to generate the adaptation transforms for the NF models. We do two passes of adaptation: in the first pass a global adaptation is performed, while the second pass uses a regression class tree. We experimented with up to 16 regression classes in the second pass, but we found that 3 classes were optimal. In this case the simulation WER improves from 43.0% to 15.4% (for the single Gaussian case). While this is a large improvement, the adapted WER, 15.4%, is still much higher than the 1.4% WER on simulated NF data (or the 1.8% WER on simulated FF data). For the case of framewise resampling, MLLR reduced the WER from 59.9% to 43.2%, again, this reduction is modest compared to the framewise resampling result for the NF case, which yielded a WER of 1.9% ; or for the FF case with matched models, which yielded a WER of 3.4%.

In short, while MLLR provides good improvements for original data, and quite substantial improvements for simulation and framewise resampling from the NF model that is recognized using FF models, the remaining errors are still substantial even for these cases that provide test data that satisfy the statistical independence assumptions.

4.1.4.2 Discriminant training via the Minimum Phone Error (MPE) approach

It is also currently standard in large speech recognition systems to incorporate discriminant model training such as MPE to reduce WER beyond what has been obtained with Maximum Likelihood (ML) models. While this approach is motivated by the desire to more effectively discriminate between correct and nearby incorrect explanations of the data, another perspective is that MPE somehow partially compensates for the dependence in the data. This is suggested by our results with MPE on our meeting data. MPE provides no improvement for the simulated or framewise resampled near-field data, for which the conditional independence assumption is satisfied; in particular, for the framewise resampled case, retraining with MPE doesn't decrease the error rate (from actually slightly increasing from 1.86% to 2.06% for the 1-Gaussian models, and staying the same at .70% for the 8-Gaussian models). For the matched far-field data case, the error rate actually increases after applying MPE, going from 3.42% to 7.10% for the 1-Gaussian

models, and from 1.32% to 1.50% for the 8-Gaussian models. In both cases, MPE provides the anticipated improvements for the original meeting data. For near-field data, MPE reduces the error rate from 44.70% to 39.00% for 1-Gaussian models, and from 33.10% to 30.90% for 8-Gaussian models. For far-field data, MPE reduces the error rate from 71.40% to 67.50% for 1-Gaussian models, and from 63.20% to 61.60% for 8-Gaussian models.

For the mismatched case, similar trends are observed. For data generated by simulation using near-field models, MPE actually makes things worse, increasing WER from 43.04% to 69.60% for the 1-Gaussian models, and from 24.80% to 34.85% for the 8-Gaussian models. For framewise resampling, WER stays roughly the same for MPE as it had been for ML models, with WER only moving from 59.93% to 59.50% for the 1-Gaussian case, and from 24.90% to 24.75% for the 8-Gaussian case. As with the matched cases, MPE does help for the original data, bringing the 1-Gaussian WER down from 84.7% to 81.9%, and the 8-Gaussian case down from 80.6% to 77.0%. These results suggest that the gain from using MPE is associated with somehow compensating for the conditional dependence in the data, since such gains are not observed when this dependence is artificially removed.

The full set of results is given in Appendix A, with contrast to the maximum likelihood results.

4.1.4.3 Discriminant features via MLP training

ICSI has been a leader for many years in MLP processing of speech to improve acoustic processing. For a number of tasks in which we used MLP outputs (after log and PCA transformations) as additional features for HMM/GMM systems, we observed significant gains. However, in general these were for tasks in which the training was reasonably representative of the test set. In our MLP experiments within this study, we found similar effects. For the near-field data, transforming the MFCC front end with a phonetically discriminantly trained MLP provided relative improvements (for the 1 Gaussian case) for the simulation, all levels of resampled data, and even (modestly) for the original data; e.g., WER dropped from 1.9% to 1.0% for the framewise resampling, and from 44.7% to 42.3% for the original data, using 1-Gaussian models. For the far-field data, similar effects were seen, although there was no improvement for the original far-field data. In particular, transforming MFCCs with an MLP reduced WER for the framewise resampling from 3.4% to 2.8%, while for the original data the error rate actually increased slightly from 71.4% to 72.2%. In both cases and for all conditions, augmenting the MFCC frontend with the MLP-processed MFCCs improved WER further. However, for the mismatch case, neither using the MLP features alone nor using them in combination with the MFCC front end provided any relief from the increased error rates; in fact, the MLP features worsened the results. For example, for framewise resampling, the error rate increased hugely from 59.9% to 92.5%. For at least this task, the MLP training seemed to overly specialize the representation to an acoustic that was clearly mismatched with the test data.

The full set of results is given in Appendix A.

4.1.4 Commentary on the efficacy of these 3 methods

As can be seen from the results briefly described above, only MLLR provided significant relief

from the huge number of errors engendered by the mismatch in acoustic data characteristics between training and test. MPE provided modest gains for the original data, but when the issues of statistical dependence are accounted for, MPE provides no gain (i.e., for the simulation and framewise resampling cases). Transforming MFCCs with an MLP is even more disappointing, as it shows no improvement for either the original data or the simulated or framewise resampled case.

There is an obvious difference between MLLR adaptation and the other two methods; the former uses information from the test set to improve performance on that very test. This is not "cheating", as there is no use of supervisory information. But it does differ significantly from the other two methods. Both MPE and MLP feature training attempt to improve discrimination on the training data, and (at least is they are ordinarily implemented) make use of no information from the test data. On the other hand, both methods provide significant gains when used for matched training and testing. This suggests that adaptation methods for discriminant methods should be explored to see if they can provide similar or better (or complementary) gains to what is seen with MLLR. We expect to be working on adaptation methods for MLP-based feature transformation in the future.

4.2 Breadth-wise study

In this section we discuss the primary results of the community survey. In Appendix B, we provide detailed information about the characteristics of the respondents, including their professional affiliation (mostly industrial or academic, with some governmental), their age (essentially all over 40, median age in their mid-50's), their position (mostly in research or development), and their professional focus (working in a range of ASR-related topics, but with roughly half directly focused on ASR itself).

In the following subsections we focus on the responses to the six questions given in section 3.2 above. Each interview was on average 30 minutes long, which also led to many anecdotal comments. We have encapsulated some of the more common themes below, and also provide figures showing a categorization of the responses. The answers to question 1, which were about the interviewees per se, are summarized in the previous paragraph and described in detail in Appendix B.

4.2.1 Question 2: Where has the current technology failed?

The interviewees cited many failures in the current speech technology. Often the failures were closely associated with the area in which the informant had been working, but in other cases they took a more global view of the technology, and attempted to tell us under which conditions the technology delivered an unacceptable result.

Figure 5 shows the technology difficulties cited by our interviewees. Many of our informants identified the lack of robustness as a primary characteristic of speech and language technology. Many responses identified the particular characteristic of speech or language that caused this lack of robustness, such as noise, the acoustic mismatch between test and training, and variability in the speaker population. The second most frequent response was that the technology was too

complicated to use or too expensive to implement. Our interviewees often noted that in order to get an application to be usable, particularly a natural dialog application, there had to be an inordinate amount of tuning or tinkering to get it to work. They observed that the amount of work that had to be done increased the cost of the application. This, too, can be seen as a failure of manageable systems to deliver robust performance in practice.

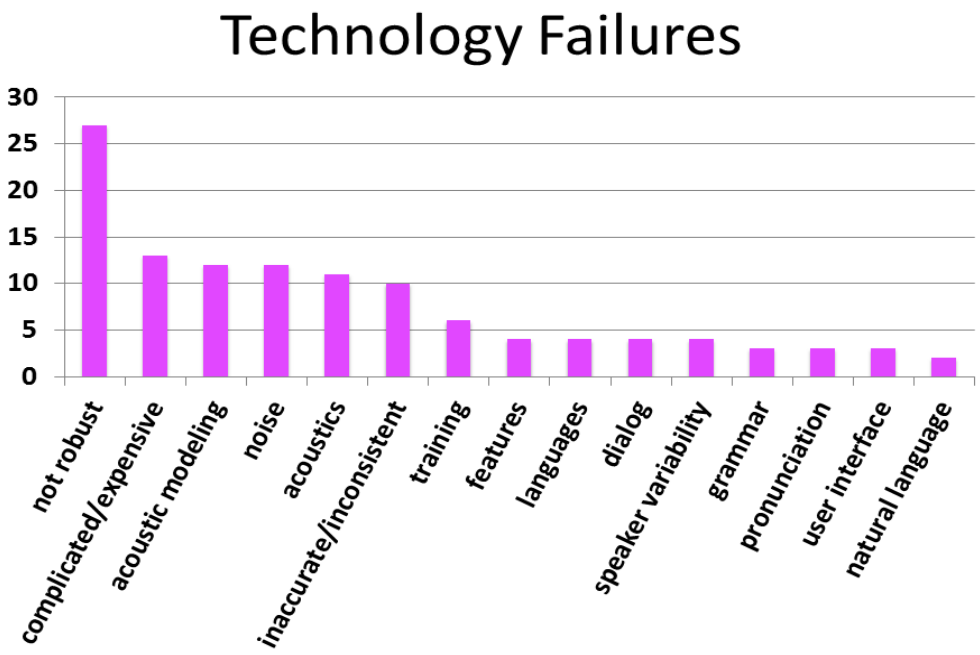

Figure 5. Categorization of responses to "Where has the current technology failed?"

It is clear that the major issue in the current applications of speech technology is the inability of current systems to perform well across different conditions. The particular conditions which were called out were performance in noise, performance in other languages, the ability to handle the variability in actual speaker populations, and general performance in acoustic conditions which differed from the training conditions. Some informants complained about the accuracy or consistency of the process, but that was a relatively infrequent response.

Representative responses we received included:

> "It's not robust to acoustic environments, multiple sources."
>
> "It fails for any conditions not seen in training, either environments or contexts."
>
> "Models are tuned too finely. Features are wrong for the job, and training is wrong."
>
> "The technology is ill equipped to handle data outside the training scenario."

There was substantial agreement that systems were too complicated or expensive, and we believe that this is simply the result of the ASR systems inability to perform robustly with simple models in the current technology. Some of the comments we received included:

> "It's not accurate enough, but to improve accuracy, or add a language or domain costs hundreds of thousands of dollars."
>
> "It requires excessive training to get adequate performance."
>
> "Pricing has impeded growth."

In short, this first technology question identified the inability of systems to generalize to noise, speaker, acoustic condition, and language as the primary problem. Current technology is brittle in a way that impedes widespread use.

4.2.2 Question 3: What do you think is broken?

This technology question was an attempt to elicit the specific cause of the technology failures noted in the previous question. While we were hoping for specific indications of technical areas that were not performing well, the question allowed for broad assessment of the problems in ASR technology.

As seen in Figure 6, there was not a consensus on what part of the technology was failing to deliver. The language model and the acoustic models were identified most frequently, but that was to be expected as, these are two of the basic building blocks of any speech recognition system. The features (or the signal processing system) followed more general complaints of lack of robustness and systems being too complicated. Less frequently cited, but still with substantial comments, were problems with adaptation and pronunciation. These were followed by more global issues of technical environmental awareness, system integration, small data, and the lack of funding for research.

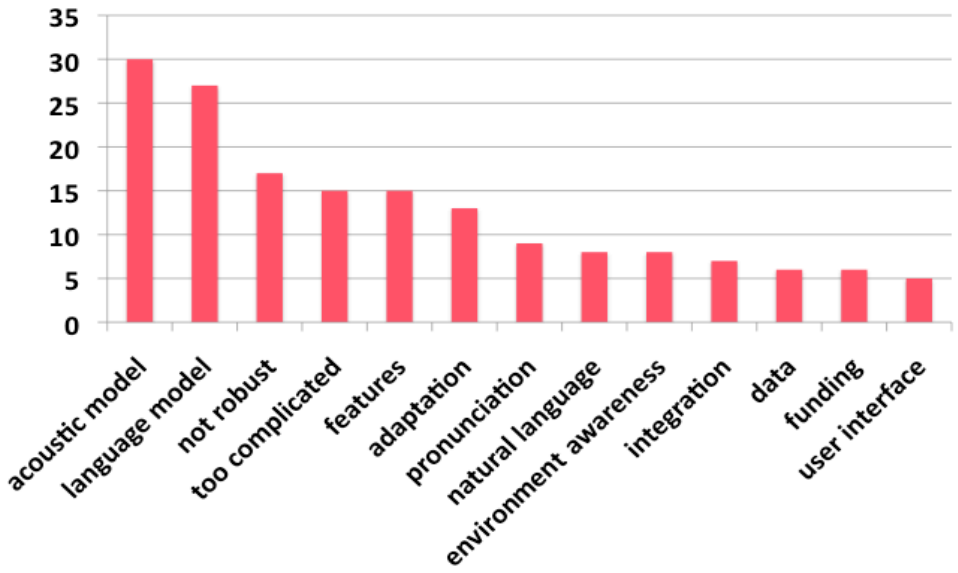

Figure 6. Categorized responses to question, "What is broken?"

Some of the comments we received included the following:

> "Acoustic models don't communicate well with language models."
>
> "We are using old models with new computational abilities; the systems are non-robust."
>
> "Most signal processing development was done in the 80's with close mic, and not using the devices we use today. There are no new models. HMMs and Cepstral analysis are still here."
>
> "The core isn't robust, and it doesn't do a good job of modeling human conversation. It assumes regimented turn-taking."
>
> "The core engines aren't robust, so we tweak as many parameters as we can, but the caller is an unwilling participant."

Far down the list was the problem of matching the user interface to the capabilities of the technology. However, while not cited frequently as one of the top two core issues with speech, this issue was another recurrent theme throughout the interview process. It was often noted that the industry (researchers, vendors, press, analysts, etc.) has oversold the capabilities of speech. While not a technological problem, it is an underlying industry problem, which leads to less adoption, acceptance, and revenue generated from speech applications. This in turn drives the perception that speech is a less valuable area to invest in.

The issue is that the industry claims that speech works well, which implies that it is easy to use. While some of the perceptual issues are caused by the deployment of speech applications that do not follow best practices in VUI design, this is only one symptom of the larger issues we uncovered in this survey. Some responses we received included:

"Its capabilities have been oversold. There are misunderstood constraints and limitations. Speech is also used inappropriately."

"It took too long to get to natural language as we have it today, but we have also over-marketed the capabilities."

In short, every major subsystem of the current ASR technology was identified by at least some interviewees as being broken. Not a single informant told us that his or her applications were successfully served by the current technology.

4.2.3 Question 4: What have you tried to do to improve the technology that should have worked but did not?

In this technology question, we attempted to assess the mental model of the users in terms of how they understood the performance of the ASR technology. Of particular importance was understanding whether the part of the technology that they pinpointed as failing was due to being difficult to mediate, or was it incorrect in some other more serious way?

The interviewees tended to think about this question more than any other. Here are some prominent examples of the replies:

"The model doesn't match the data."

"Pronunciation modeling, acoustic modeling, and scaling in the language models didn't work.

"Pronunciation modeling has failed for us. We have worked hard for very little payback."

"Noise - new algorithms aren't good enough; accent models need to be broader."

"I worked on ASR for people and place names. One project I did grammars for every possible pronunciation and it slowed the recognizer down too much."

"We tried to get more data from our domains to get different accents. After hundreds of hours of data, there was very little improvement at all."

"Predicting user reaction to a prompt failed. We think we have a prompt nailed, and feel it's intuitive, but in the real world it isn't."

"Tried emotion detection."

"Auditory representations haven't helped much. Brute force techniques need too much data, and it is difficult to incorporate NLP."

"Microphone related projects - impossible to predict performance from data."

"I tried to model different parts of the sentence differently. For example, we gave information at the beginning and content at the end, with the verb as the pivot point. But it didn't improve anything."

"I tried to model non-linear acoustics."

Figure 7 shows the categories of what interviewees tried to fix in speech that should have worked but did not. The answers were reasonably in agreement. While "fixing" training, adaptation, or features were standard portions of ASR system efforts, it is striking that every respondent who tried to adjust pronunciation failed to make his or her systems significantly better. Attempts to substantially enhance performance with emotions or more sophisticated grammars generally failed as well, as did attempts to make the systems more sophisticated by redesign. While some people have had success using posterior probabilities rather than feature measurements in the process, the success was not universal. Note that major improvements through adaptation were done decades ago in the form of VTLN or MLLR, and these early successes have been difficult to extend further.

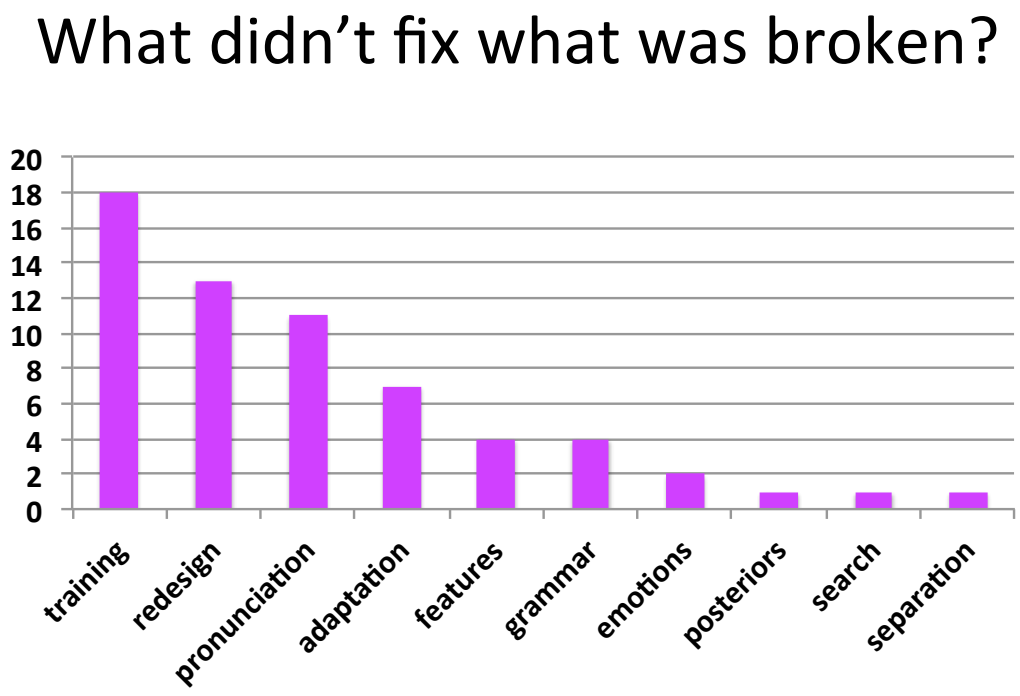

Figure 7. Categorization of unsuccessful attempts to fix the technology

4.2.4 Question 5: Why didn't your fix work?

The question of why did it fail was meant to elicit the reasons for the lack of success cited in the previous question. Again, answers were mixed between specific technical issues that were attempted and more general comments about the speech recognition technology itself.

Figure 8 shows the answers given for "Why didn't your fix work?" Respondents were generally in agreement that the technology was not mature. Several of them said this directly, while many others complained about the lack of standardization, the immaturity of the particular models

(especially the language model), having incorrect training data that did not match the speech recognition task data, and the unpredictability of rare events and of speech in general. We have lumped these all together under "immature technology".

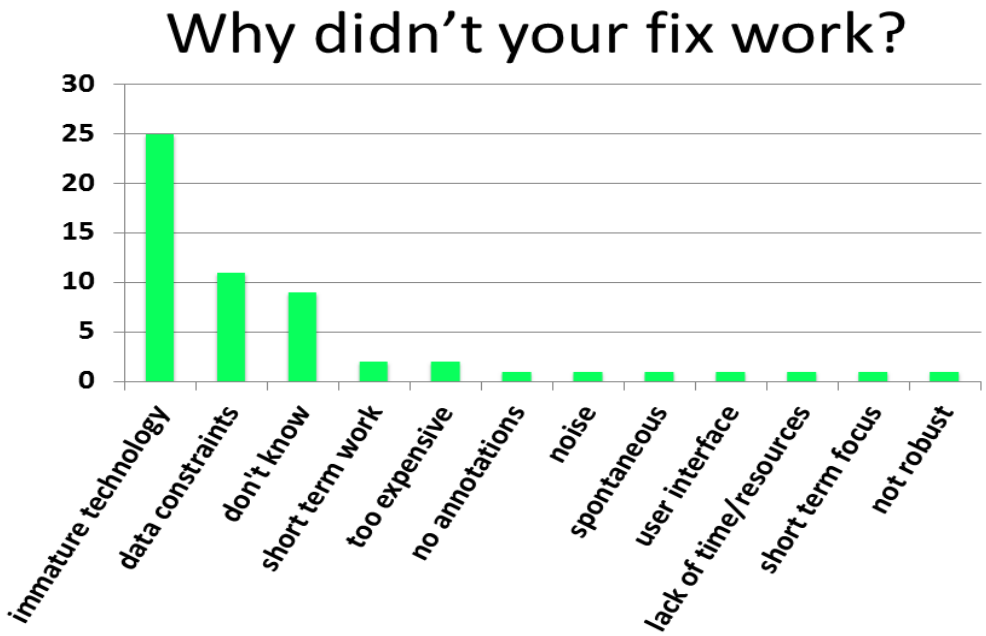

Figure 8. Categorizations of responses explaining why their plausible solutions to ASR technology limitations had not succeeded

Some of the comments we received as to the technology being immature included:

> "Combination of noise and spontaneous conversation"
>
> "Don't know. It worked somewhat in lab, but not live"
>
> "It was much more complex than originally thought, and there wasn't enough data".
>
> "There wasn't enough training data to anticipate reaction, and the way people respond can change based on world, changes and other factors."
>
> "People are unpredictable, and real world ASR doesn't understand that. ASR is dumb."
>
> "We couldn't get enough data."
>
> "The current models aren't tuned to spontaneous speech, and don't take into account semantic and syntactic info."

We received a number of comments on less problematic issues as well. Some interviewees said

that a serious issue was not having enough data. For example, when creating grammars for new languages, data was often sparse and restricted to a few speakers, and online text was not available. Others cited the short-term focus of research, and a small but significant proportion of the respondents were unable to specify the problems in detail. A few noted specific situations in which their attempts failed, such as with noisy or spontaneous speech.

4.2.5 Question 6: Have you solved any speech technology problems that were not published? If so, what?

The general answer to this question was "no", and there were two reasons for the no's. The first reason was that the respondent had done all of the work under government sponsorship or at an institution that made everything public. In this case, everything was essentially made available in some form, so there were no hidden solutions. The second reason for saying no came from our corporate interviewees, who said they weren't allowed to say because the results were either trade secrets or in patents that were pending.

There were some pointers to old work that would not be relevant to the current issues. For example, a few respondents spoke of integration issues, and a slightly larger number of people noted that their solutions were "simply" engineering solutions and not generally applicable to the larger technology.

Several people told us of things that they had fixed, but didn't publish because the projects ran out of funds, or that their work was accomplished in conjunction with other work that didn't merit a separate research paper. Despite our hope that we would discover a hidden mine of essential but unshared technical gold, we were disappointed.

4.2.6 Summary of responses to technical questions

Our interviewees believed that they were working with an emerging technology. In fact, there was a note of cynicism from many as they felt that the core recognition models were so old, that the technology had been an emerging technology for 30 years. It was described as immature in essentially all of the technical aspects of recognition. While there was minor dissatisfaction with recognition performance per se, the major complaint was that the speech recognition systems that are deployed today are not robust to conditions other than the training conditions. They degrade rapidly and not gracefully in noise, for novel speakers, in far-field or other unusual acoustic conditions, in accented speech, and for speech in which other signals or noises share the acoustic channel.

Our respondents identified essentially every element of the current ASR technology as the focus of experiments to attempt to improve the technology. Failures were abundant, and performance continues to lag that of people in similar situations.

Our industry poll suggests that a critical issue with the current speech technology is that it is not robust to variability that is transparent to human listeners. That is, our artificial systems degrade much more quickly than human listeners for acoustic situations unlike those in the training material, accents or non-standard use of grammatical constructions, noise or reverberation, and all types of interfering signals. One other common thread is that performance of current systems is difficult to predict for any particular acoustic signal.

4.2.7 A historical and literature perspective

The "standard" historical chart for speech recognition performance cited by many in the field is a NIST historical performance chart. It originally was created to track DARPA funded government programs, and has recently been extended to cover some non-DARPA programs. The chart may be found on the NIST website, and the most recent version dates from May, 2009.

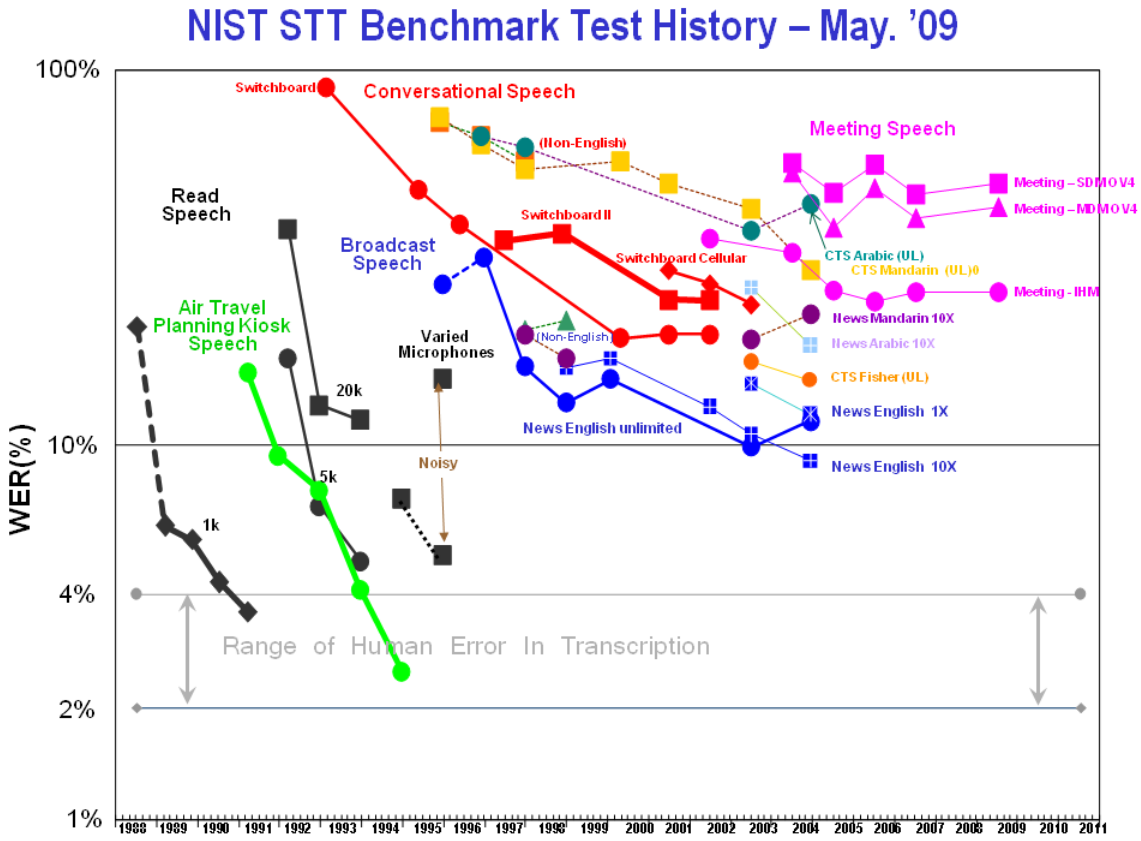

Figure 9. This shows the official NIST ASR History graph, found at http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html

Figure 9 summarizes system accuracy in terms of word error rate for the NIST evaluated programs over the last 20 years. Note the distinct flattening in performance over the last decade for sophisticated systems attempting difficult tasks such as conversational speech or meeting speech. Of particular concern is the lack of progress in meeting speech and in conversational speech in the past decade. Note that the U.S. government has not funded ASR foundational work over those recent years, but has focused on particular implementations and narrower data

regimes, or ASR as a component in a larger task (such as machine translation form speech). The meeting speech evaluations primarily took advantage of European funding of several groups.

The NIST chart should not be taken to denigrate the work of many research teams who have each made incremental progress on the particular systems. It does indicate, however, the difficulty in finding robust technological advances with wide applicability.

In the past two decades, there have been at least six surveys of speech and language technology that attempted to report the state of the art; the latest of which was actually a 2012 special issue of the IEEE Signal Processing Magazine that itself contained 6 relevant articles. We attempt here to summarize the major findings of these surveys, and to assess the similarity of their findings to our industry poll.

4.2.7.1 As found in earlier surveys of speech recognition

a. In the introduction to [24], Furui, Deng, Gales, Ney, and Tokuda note, "Despite the commercial success and widespread adaptation, the problem of LVCSR is far from being solved; background noise, channel distortion, foreign accent, and casual and disfluent speech or unexpected topic change can cause automated systems to make egregious recognition errors. This is because current LVCSR systems are not robust to mismatched training and test conditions and cannot handle context as well as human listeners, despite being trained on thousands of hours of speech and billions of words of text". We note that this is essentially what our informants also told us.

The following 6 paragraphs (b 1-6) give a few key points from each of six papers within this special issue.

b.1 [25], the first of several relevant papers in the recent survey volume, is focused on speech recognition systems associated with government programs. The authors note advances in front ends, speaker adaptation, acoustic modeling, discriminative training, noise adaptation, and more sophisticated language modeling. However, they reiterate that LVCSR is far from being solved.

b.2. In the second paper in this volume, [26], the authors note that front ends which mimic biological or psychoacoustic properties "have in many cases provided significant reductions in errors, and we are experiencing a resurgence in community interest." They note RASTA and cepstral mean subtraction as proving significant improvement. (RASTA was developed in 1991 [27], and cepstral mean subtraction dates from 1981 [28]. The authors do not report recent improvements, but rather report renewed interest in this area in hope of increasing the robustness of speech recognition systems.

The authors note further that "While machines struggle to cope with even modest amounts of acoustic variability, human beings can recognize speech remarkably well in similar conditions: a solution to the difficult problem of environmental robustness does indeed exist. While a number of fundamental attributes of auditory processing remain poorly understood, there are many instances in which analysis of psychoacoustic or physiological data can inspire signal processing research" (page 36). It appears that the solution in practice remains elusive.

b.3. In the third paper of the special issue, [29], the authors recount experiments with an alternative to the standard phonetic representations. They, like our informants, find a brittle system; "That speech recognition is worse for both conversational and hyper-clear speech suggest that the representation used in today's recognizers may be flawed" (page 46). While recounting in detail a subword alternative, they do not find substantial performance increases using these methods.

b.4. In the fourth paper of the review, [30], the authors review many of the models for discriminative training of modern recognizers)[5]. In this paper, the authors recount many heuristics, including Minimum Bayes Risk, and Margin Based Training, showing gains of 10-20 percent over ML systems. But these improvements are not new, and they do not improve the robust performance problem. In fact, the authors note that "it might be worth rethinking computing models and considering alternative architectures" (Page 68).

b.5. The fifth paper, [31], discusses more sophisticated models for discriminative training. However, the authors do not offer performance measures, and they do note that "Though current state-of-the-art systems yield satisfactory recognition rates in some domains, performance is generally not good enough for speech applications to become ubiquitous" (Page 71).

b.6. The sixth paper, [32], recounts the newest "big thing" in ASR is using many-layer nets, despite the fact that "DNNs with many hidden layers are hard to optimize". The authors recount a large number of examples of speech recognition in which DNN systems perform better than "good" modern systems. The comparisons, while interesting, do not compare "best" modern systems with DNNs, and thus simply set the stage for more work. It does appear that DNN systems are efficient at training from limited data, but the heuristic nature of the solutions left the situation in doubt. In any case, these solutions were not assessed for their ability to generalize to unseen data, noisy conditions, or other novel situations.

In summary, the latest substantive review of the state-of-the-art in speech recognition finds a number of key flaws in the current technology. It can be used in some circumstances, but there is not a clear direction forward, except for "more work".

Here are five other reviews of the state of the art in speech recognition that are also relevant:

c. In a two-part article published in 2009 [33], the authors offer a view of ASR technology not dissimilar to the 2012 papers cited above. The authors cite identical "advances", most of which occurred a decade or more before the review. They note, "The most significant paradigm shift for speech-recognition progress has been the introduction of statistical methods, especially stochastic processing with hidden Markov models (HMMs) in the early 1970s. More than 30 years later, this methodology still predominates. Statistical discriminative training techniques are

[5] The original HMM formulation, using Maximum Likelihood (ML) training, can be shown to converge to an error minimum (in the presence of infinite training data) only in the case that the data were generated by the same geometry as the model. For speech recognition this most basic constraint does not hold, and discriminative training has been used since the '80s to improve performance of models trained on the ML criterion. (One author was a member of the IBM speech recognition group at Yorktown in 1984, where an initial implementation of discriminative training was used to improve the performance of Tangora, the early 5000-word office dictation system).

typically based on utilizing maximum mutual information (MMI) and the minimum-error model parameters. Adaptation is vital to accommodating a wide range of variable conditions for the channel, environment, speaker, vocabulary, topic domain, and so on."

Despite all of this progress, the authors cite "grand challenges" which remain. Of those, the first is dealing with everyday audio. They note, "This is a term that represents a wide range of speech, speaker, channel, and environmental conditions that people typically encounter and routinely adapt to in responding and recognizing speech signals. Currently, ASR systems deliver significantly degraded performance when they encounter audio signals that differ from the limited conditions under which they were originally developed and trained". The authors further suggest challenges of self-adaptive language, rapid portability, detection of rare events, and others. These comments are consistent with those of our interviewees.

d. The view from across the ocean is the same. In [34], the authors write "The interaction between a human and a computer, which is similar to the interaction between humans, is one of the most important and difficult problems of the artificial intelligence. Existing models of speech recognition yield to human speech capabilities yet; it evidences of their insufficient adequacy and limits the introduction of speech technologies in industry and everyday life". In other words, the poorer performance of speech recognition systems limits their use by the population at large.

e. In [35], the authors note: "In most speech recognition tasks, human subjects produce one to two orders of magnitude less errors than machines. There is now increasing interest in finding ways to bridge such a performance gap. What we know about human speech processing is very limited."

f. A review that reads as remarkably modern (despite being 13 years old), [36], was written from a European perspective. They authors said, "Most of today's state-of-the-art systems for transcription of broadcast data employ the techniques described in Section II, such as PLP features with cepstral mean and variance normalization, VTLN, unsupervised MLLR, decision tree state tying, and gender- and bandwidth-specific acoustic models. Over the past four years, tremendous progress has been made on transcription of broadcast data. State-of-the-art transcription systems achieve word error rates around 20% on unrestricted broadcast news data, with a word error of about 15% obtained on the recent NIST test sets. … Despite the numerous advances made over the past decade, speech recognition is far from a solved problem, as evidenced by the large gap between machine and human performance. The performance difference is a factor of five to ten, depending upon the transcription task and test conditions."

This analysis could have been written today – the basic observations still hold, and if anything the laboratory-measured error rates Gauvain and Lamel cite are optimistic!

g. It is particularly sobering to read Richard Lippmann's review of speech recognition from 1997 [37]. He states "Error rates of machines are often more than an order of magnitude greater than those of humans for quiet, wideband, read speech. Machine performance degrades further below that of humans in noise, with channel variability, and for spontaneous speech." This comment could be made today, although there are now a few situations where human performance is approximated in narrow domains (see below).

It is also interesting to look at outlying experiments in the use of the current speech recognition technology. Two in particular are of interest:

h. In a recent (2012) Google tech report [38], the authors note that recognition performance falls to 17% search term error for language models built on 230 billion words of text. Informal discussions suggest that the acoustic models used are trained on centuries of speech. In short, much more data may help but it does not fix the performance issue for current systems.

On the other hand, during the GALE project, SRI (and other participants) demonstrated 2.4% character error rates for the recognition of Mandarin broadcast news. After substantial analysis SRI researchers learned that Mandarin broadcast speakers in China are schooled in the same accent and, further, are taught to talk in a standard cadence. Thus, Mandarin broadcast news has a regularity unseen in other speech, but which the hidden Markov models capitalize on for superior performance. [39].

Some other relevant documents are listed in Appendix C.

4.2.7.2 Commentary on the literature survey

As is obvious from these papers and articles, the performance deficiency of current speech technology compared to human performance (noted by Lippman in 1997) is still observed for current applications of speech technology. While several techniques have been developed for more advanced acoustic observations, adaptation, language model smoothing, and vocal tract length modeling, the rate of decrease of error rates over time has slowed drastically over the past decade. (Of note is the fact that there has been no U.S. government funding of basic research or engineering in speech over that same decade). Each review, and many of the papers citing better performance in particular circumstances, notes that our recognitions systems are not robust to noise, reverberation, different speakers, and accent, and that they are too complicated to port easily to new circumstances or to new languages. In short, the speech recognition field has developed a collection of limited solutions to constrained speech problems, and these solutions fail in many situations in the world at large. Their failure modes are acute but unpredictable and non-intuitive, thus leaving the technology defective in broad applications, and difficult to manage even in well behaved environments.

5.0 CONCLUSIONS

The state of speech recognition, as reported by our interviewees, is awaiting a transition from a difficult, immature technology to a robust, mature technical system[6]. Our interviewees identified every portion of the current technology as defective, and in turn identified those same areas as the focus of work that has failed to fix the major problems. The major issue seems to be the lack of robust performance, leading to system failures for acoustic and linguistic variabilities that do not bother human listeners. This failure makes system design difficult, as our systems break in unpredictable and unintuitive ways. A secondary problem is that these systems do not perform well for sophisticated tasks like spontaneous dictation, although that may be a problem with more than the non-robust performance problem.

[6] As one of us has noted in a recent presentation, one can be old and still be immature.

The survey of practitioners of speech and language technology reported here finds that modern speech recognition systems are brittle, non-robust, and overly complex. The systems fail to generalize outside the domain of the training data, and within the training domain they fail for moderately complicated tasks such as meeting transcriptions.

Every aspect of the speech recognition technology has been exercised in an effort to make the performance better and more robust. Despite several decades of small incremental improvements in performance (nearly all of which occurred prior to the last decade) overall performance appears to have plateaued. A survey of the literature in speech recognition confirms the continuing inability of our systems to mimic human performance in the presence of noise, reverberation, different dialects, different languages, and other variations which are part of the everyday environment.

In the past decade, researchers have greatly increased the size of our datasets, the number of free parameters in our models, and the amount of computation available for both training and testing of our systems. While performance has been improved along many dimensions, the final result is qualitatively the same as those of a decade ago.

For the specific issue of the acoustic model, by exploiting the method of resampling, we constructed a series of pseudo datasets from near-field and far-field meeting room datasets. The most artificial of these satisfied the HMM model assumptions, while at the other extreme, the resampled data deviated from the model in the way real data did. Using these datasets we probed the standard HMM/GMM framework for automatic speech recognition. Our results showed that when the conditions are matched (even if they are far-field), the model errors (i.e., errors due the incorrect assumption of conditional independence) dominate; however, in mismatched conditions, the standard ASR features computed from far-field data are neither invariant nor separable with near-field models, and contribute significantly to the total errors; these basic conclusions are illustrated in Figures 10 and 11. We then studied unsupervised MLLR adaptation and MPE training as the means to compensate for this issue in the model space; while these approaches mitigate the errors somewhat, the conclusions about the lack of invariance of the MFCC features in varying acoustic conditions still holds true. Finally, we also used discriminatively trained MLPs to transform the MFCCs, and these too failed to alter the conclusion about MFCCs. On the contrary, the highly discriminant MLP training actually worsened performance for all the experiments under the mismatched condition.

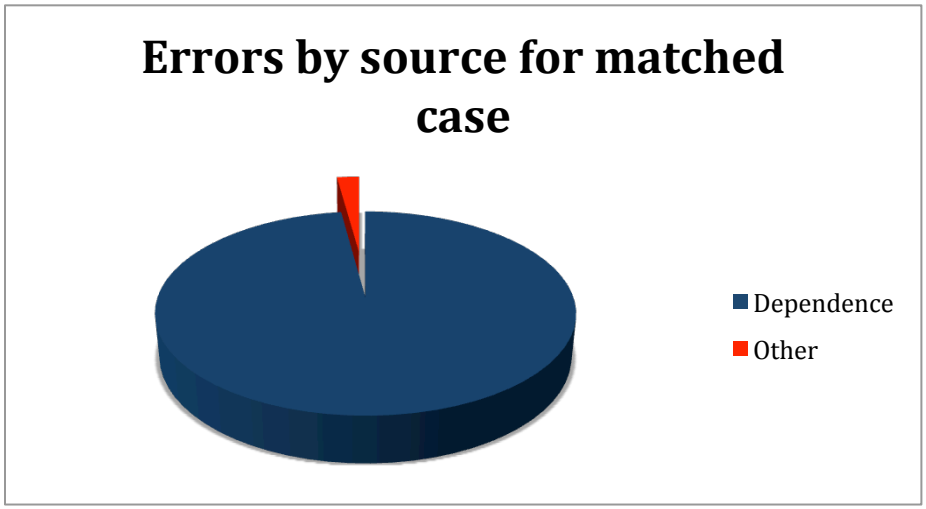

Figure 10. Inferred error proportions for sources of word errors in recognition of near-field meeting data from models trained on near-field data, ICSI Meeting Corpus. "Dependence" refers to the conditional independence assumptions common to HMMs. "Other" includes all other sources of error (LM, front end deficiencies, pronunciation models, etc.). This figure refers to the 8-Gaussian models and the error rates given in Appendix A.

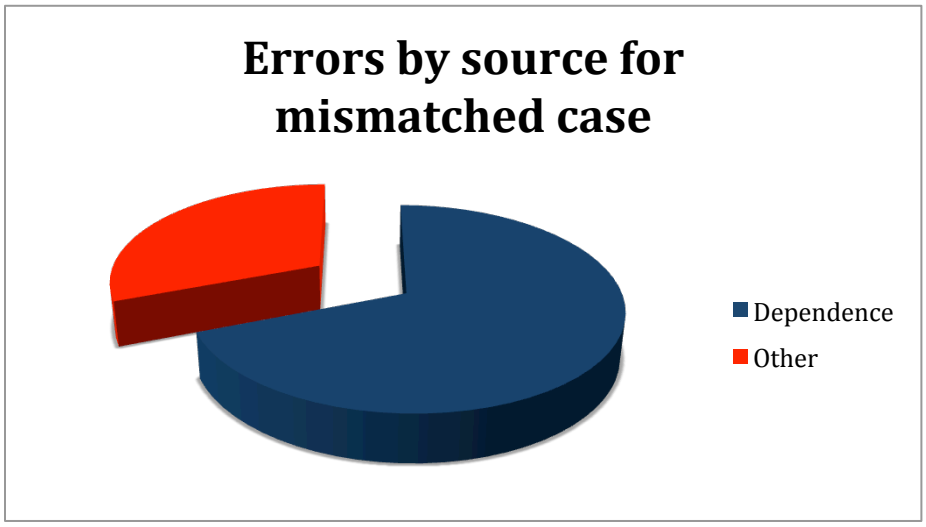

Figure 11. Inferred error proportions for sources of word errors in recognition of far-field meeting data from models trained on near-field data, ICSI Meeting Corpus. "Dependence" refers to the conditional independence assumptions common to HMMs. "Other" includes all other sources of error (LM, front end deficiencies, pronunciation models, etc.), but the LM and pronunciation models are presumed to be as good here as for the matched case; the primary difference is likely to be acoustic, so the front end is the likely suspect. This figure refers to the 8-Gaussian models and the error rates given in Appendix A.

6.0 RECOMMENDATIONS

6.1 Make use of diagnostic analysis to drive development of remedies

It has been suspected for some time that, for instance, the inaccuracy of the standard conditional independence assumption in the acoustic model is a key reason for the high error rates in fluent speech recognition; further, it has been largely assumed that the lack of invariance of ASR signal processing to variability in acoustic conditions is also a significant source of errors. Our study has confirmed both of these points. However, there is potential for much greater gain than "simply" confirming our preconceptions. As researchers propose potential remedies, there is now a method for analyzing the effects of their proposed methods with greater specificity and utility than simply seeing if the word error rate went down. For instance, while segment models and episodic approaches both might be able to better handle local statistical dependence, the details probably matter – and using methods such as those that we developed here could be useful in a host of decisions made in the development of alternative methods.

6.2 Extend diagnostic analysis to other components

We have shown that the independence assumption in acoustic modeling, particularly for frames within a state, is a significant remaining problem, even under matched acoustic conditions. Is there a related problem with other components, such as the language model? Much as with the acoustic model, attempts to transcend the limitations of the simple n-gram have yielded only incremental improvements. It is likely that moving beyond this point will not be possible without effective diagnostics, ones that are more specific than word error rate (or certainly more effective than perplexity).

6.3 Update the model

This would be an opportune time to reconsider the decades-old HMM formulation, and search for models that better capture speech and language characteristics. We should enlist the help of theorists (such as those who will be associated with the new Simons Center for Theoretical Computer Science at Berkeley) to derive a better model. Whoever studies the problem should have a particular focus on the case of mismatch between training and test data, i.e., on generalization.

6.4 Seek low dimensional parameters to characterize speech variability

Many phenomena arise from complex interactions between many components; the production and perception of speech by the human brain is an example of such phenomena. Consequently, it may be the case that the recognition of speech is and must be complicated. On the other hand, some of the cases of significant progress in ASR (e.g., VTLN, RASTA, cepstral mean subtraction) are surprisingly simple. Consequently, it would be worthwhile to seek to develop systems that automatically account for predictable variations from the training data without specific training for that condition, where the obvious conditions one would like to compensate for are far-field acoustics, additive noise, speakers with light accents or dialects, and informal spontaneous speech.

6.5 Study the brain

There is an existing significant example of speech recognition that actually works well in many adverse conditions, namely, the recognition performed by the human ear and brain. Methods for analyzing functional brain activity have become more sophisticated in recent years, so there are new opportunities for the development of models that better track the desirable properties of human speech perception. While many such methods have been tried before and have provided, at best, limited improvements, recent improvements in basic brain scan technology (e.g., "eCog", which collects data directly from the surface of the human cortex) provides an opportunity to significantly limit the vast search space of all possible ASR approaches. In particular, this field of knowledge should be mined to assist in the design of new acoustic front ends that would be more invariant to signal variability that is independent of the linguistic content.

6.6 Beyond ASR

The "in-depth" study described in this report was focused specifically on speech recognition. That being said, since the use of HMMs has spread far beyond speech processing, there are many fields of inquiry that are also limited by the limitations of these models. One application of the methods described here to many other fields, e.g. speech synthesis, machine translation, part of speech tagging, bioinformatics (e.g., DNA sequencing, gene prediction), protein folding, and time series analysis.

More generally speaking, HMMs are a staple of machine learning as applied to many tasks requiring the decoding of sequences, and there are likely improvements that could be found in many areas given improved diagnostic methodology. In speech recognition research, very little diagnostic analysis has ever been undertaken, and we would argue that as a result progress in the field has proceeded largely by trial and error and it has been susceptible to fads (success of an interesting technique in a very different field leads to "trying it out" in speech recognition: wavelets, compressed sensing, deep learning, etc.). In the more general field of machine learning, very little effort has been expended on understanding how algorithms fail when applied to real world problems outside the laboratory. We anticipate that encouraging more of a diagnostic spirit for machine learning research could have very broad effects, much as the introduction of HMMs to this field did earlier.

7.0 REFERENCES

[1] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, Vol. 7, No. 1, pp. 1–26, 1979.

[2] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Gelbart, D. Ellis, G. Doddington, B. Chen, B., O. Cetin, H. Bourlard, and M. Athineos. Pushing the envelope aside: Beyond the spectral envelope as the fundamental representation for speech recognition. *Signal Processing Magazine*, IEEE, Vol. 22, No. 5, pp. 81 –88, 2005.

[3] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke. Understanding and improving speech recognition performance through the use of diagnostic tools. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995), 1995.

[4] L. Chase. Error-Responsive Feedback Mechanisms for Speech Recognizers Ph.D. thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, April 1997.

[5] G. Saon and J.-T. Chien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. *Signal Processing Magazine*, IEEE, Vol. 29, No. 6, pp. 18 –33, November 2012.

[6] G. Heigold, H. Ney, R. Schluter, and S. Wiesler. Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance. *Signal Processing Magazine*, IEEE, Vol. 29, No. 6, pp. 58 –69, November 2012.

[7] M. Gales, S.Watanabe, and E. Fosler-Lussier. Structured discriminative models for speech recognition: An overview. *Signal Processing Magazine*, IEEE, Vol. 29, No. 6, pp. 70 –81, November 2012.

[8] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *Signal Processing Magazine*, IEEE, Vol. 29, No. 6, pp. 114 –126, November 2012.

[9] R.M. Stern and N. Morgan. Hearing is believing: Biologically inspired methods for robust automatic speech recognition. *Signal Processing Magazine*, IEEE, Vol. 29, No. 6, pp. 34 –43, November 2012.

[10] S. Wegmann and L. Gillick. Why has (reasonably accurate) automatic speech recognition been so hard to achieve?" arXiv:1003.0206 [cs.CL], 2010.

[11] D. Gillick, L. Gillick, and S. Wegmann. Don't Multiply Lightly: Quantifying Problems with the Acoustic Model Assumptions in Speech Recognition. Proceedings of the 2011 IEEE Automatic Speech Recognition and understanding Workshop (ASRU 2011), pp. 71–76, 2011.

[12] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus." Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), 2003.

[13] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng. The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System. Proceedings of the Second International Workshop on Classification of Events, Activities, and Relationships (CLEAR 2007) and the Fifth Rich Transcription 2007 Meeting Recognition (RT 2007), 2007.

[14] RT-2002 Evaluation Plan.
http://www.itl.nist.gov/iad/mig/tests/rt/2002/docs/rt02_eval_plan_v3.pdf

[15] The RT-04S Evaluation Data Documentation.
http://www.itl.nist.gov/iad/mig/tests/rt/2004-spring/eval/docs.html

[16] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun. The Rich Transcription 2005 Spring Meeting Evaluation.
http://www.itl.nist.gov/iad/mig/tests/rt/2005-spring/index.html

[17] Dan Ellis. SKEWVIEW – Tool to visualize timing skew between files. 2011.
http://labrosa.ee.columbia.edu/projects/skewview/

[18] S.Y. Chang. ICSI Meeting Alignments. 2012.
http://www1.icsi.berkeley.edu/~shuoyiin/research/meetingskew/chanskew.html

[19] S.J. Young, G. Evermann, M.J.F. Gales, D. Kershaw, G. Moore, J.J. Odell, D.G. Ollason, D. Povey, V. Valtchev, and P.C. Woodland. *The HTK Book*, Version 3.4, 2006.

[20] O. Cetin and A. Stolcke. Language modeling in the ICSI-SRI Spring 2005 meeting speech recognition evaluation system, International Computer Science Institute Technical Report TR-05-006, Berkeley, California, July 2005.

[21] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, No. 9, 1995.

[22] J. Faugier and M. Sargeant. Sampling hard to reach populations. *Journal of Advanced Nursing*, Vol. 26, No. 4, pp. 790-797, October 1997.

[23] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret. Incremental on-line feature space MLLR adaptation for telephony speech recognition. Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002), Denver, Colorado, pp. 1417-1420, September 2002.

[24] S. Furui, L. Deng, M. Gales, H. Ney, and K. Tokuda. Fundamental Technologies in Modern Speech Recognition. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, November 2012.

[25] G. Saon and J.-T. Chien. Large-Vocabulary Continuous Speech Recognition Systems. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 18-33, November 2012.

[26] R.M. Stern and N. Morgan. Hearing is Believing: Biologically Inspired Methods for Robust Automatic Speech Recognition. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 34-43, November 2012.

[27] H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 578-589, October 1994.

[28] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 29, No. 2, pp. 254-272, April 1981.

[29] K. Livescu, E. Fosler-Lussier, and F. Metze. Subword Modeling for Automatic Speech Recognition: Past, Present, and Emerging Approaches. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 44-57, November 2012.

[30] G. Heigold, H. Ney, R. Schluter, and S. Wiesler. Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 58-69, November 2012.

[31] A. Ragni and M.J.F. Gales. Structured Discriminative Models for Speech Recognition. Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), Prague, Czech Republic, pp. 4788-4791, May 2011.

[32] G. Hinton, L. Deng, Y. Dong, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82-97, November 2012.

[33] J. Baker, L. Deng; J. Glass, S. Khudanpur, C. L. Lee, N. Morgan, and D. O'Shaughnessy. Developments and directions in speech recognition and understanding, Part 1 [DSP Education]. *IEEE Signal Processing Magazine*, Vol. 26, No.3, pp.75-80, May 2009.

[34] A. L. Ronzhin, R. M. Yusupov, I. V. Li, and A. B. Leontieva. Survey of Russian Speech Recognition Systems. Proceedings of the 11th International Speech and Computer Conference (SPECOM'2006), St. Petersburg, Russia, June 2006.

[35] M.A. Anusuya and S.K. Katti. Speech Recognition by Machine: A Review. *International Journal of Computer Science and Information Security*. Vol. 6, No. 3, 2009.

[36] J. Gauvain and L. Lamel. Large-Vocabulary Continuous Speech Recognition: Advances and Applications. *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1181-1200, August 2000.

[37] R. P. Lippmann. Speech recognition by machiens and humans. *Speech Communication*, Vol. 22, No. 1, pp. 1-15, July 1997.

[38] C. Chelba, D. M. Bikel, M. Shugrina, P. Nguyen, and S. Kumar. Large Scale Language Modeling in Automatic Speech Recognition. Google Technical Report, 2012.

[39] W. Wang, A. Mandal, X. Lei, A. Stolcke, and J. Zheng. Multifactor Adaptation for Mandarin Broadcast News and Conversation Speech Recognition. Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom, Vol. 9, pp. 2103-2106, September 2009.

[40] S. Sagayama, K. Shinoda, M. Nakai, and H. Shimodaira. Analytic Methods for Acoustic Model Adaptation: A Review. Proceedings of the ISCA Workshop on Adaptation Methods for Speech Recognition, Sophia Antipolis, France, pp. 67-76, August 1991.

[41] A. Mansikkaniemi. Acoustic Model and Language Model Adaptation for a Mobile Dictation Service. Master's Thesis, Aalto University School of Science and Technology, March 2010.

[42] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Garret, and B. Strope. Your Word is my Command: Google Search by Voice: A Case Study. *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, pp. 61-90, Springer Verlag, 2010.

[43] X. Xiao, J. Li, E. S. Chng, H. Li, and C.-H. Lee. A Study on the Generalization Capability of Acoustic Models for Robust Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 6, pp. 1158-1169, August 2010.

[44] L. Deng, X. Li, D. Yu, and A. Acero. A Hidden Trajectory Model with Bi-Directional Target Filtering: Cascaded vs. Integrated Implementation for Phonetic Recognition. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Vol. 1, pp. 337-340, March 2005.

[45] A. Gunawardana and A. Acero. Adapting Acoustic Models to New Domains and Conditions Using Untranscribed Data. Proceedings of the ISCA International Conference on Speech Communication and Technology, September 2003.

8.0 APPENDIX A – DETAILED NUMERICAL RESULTS, IN-DEPTH STUDY

In the two subsections that follow, we provide tables summarizing all of the major results for the simulation and resampling studies.

8.1 Maximum Likelihood and Minimum Phone Error results, simulation and resampling studies

Table 5 provides the word error rates for all conditions: matched near-field, matched far-field, and mismatched (near-field models, far-field data), for both maximum likelihood and discriminatively trained (MPE) models.

Table 5. Maximum likelihood vs MPE word error rates for the 3 conditions under study, and for 1, 2, 4, and 8 Gaussian components per crossword triphone. The μ columns give the average word error rate over the different jackknife cuts for each case, and the SE columns give the corresponding standard error measure. The left hand column of each table gives the type of each experiment, ranging from simulation from the model through the different levels of resampling, and ending in the case of recognition with the original meeting data

ML - MPE

ML

Close mic

	1g		2g		4g		8g	
	μ	SE	μ	SE	μ	SE	μ	SE
sim	**1.42**	0.03	**0.52**	0.02	**0.50**	0.01	**0.50**	0.01
frm	**1.86**	0.05	**0.86**	0.02	**0.78**	0.02	**0.70**	0.01
state	**9.60**	0.17	**6.54**	0.10	**5.04**	0.08	**4.24**	0.09
phn	**21.44**	0.21	**16.14**	0.25	**12.00**	0.26	**9.38**	0.40
word	**37.56**	0.28	**30.68**	0.37	**25.20**	0.23	**21.22**	0.24
orig.	**44.70**		**38.60**		**35.20**		**33.10**	

Far mic

	1g		2g		4g		8g	
	μ	SE	μ	SE	μ	SE	μ	SE
sim	**1.82**	0.03	**1.13**	0.02	**1.10**	0.02	**1.00**	0.02
frm	**3.42**	0.02	**1.68**	0.03	**1.40**	0.04	**1.32**	0.05
state	**23.24**	0.20	**19.14**	0.13	**16.94**	0.30	**15.22**	0.22
phn	**45.46**	0.41	**38.14**	0.24	**33.42**	0.26	**29.32**	0.19
word	**63.54**	0.45	**58.12**	0.31	**53.38**	0.19	**49.38**	0.09
orig.	**71.40**		**68.50**		**66.10**		**63.20**	

Mismatched conditions

	1g		2g		4g		8g	
	μ	SE	μ	SE	μ	SE	μ	SE
sim	**43.04**	0.23	**34.01**	0.17	**27.20**	0.14	**24.80**	0.25
frm	**59.93**	0.26	**38.32**	0.15	**29.28**	0.11	**24.90**	0.15
state	**75.82**	0.27	**61.46**	0.31	**55.64**	0.40	**52.92**	0.23
phn	**80.58**	0.29	**72.76**	0.31	**68.26**	0.43	**64.44**	0.33
word	**80.36**	0.15	**77.42**	0.18	**74.64**	0.24	**71.98**	0.17
orig.	**84.70**		**83.30**		**81.90**		**80.60**	

MPE

Close mic

	1g		2g		4g		8g	
	μ	SE	μ	SE	μ	SE	μ	SE
sim	**1.57**	0.02	**0.68**	0.03	**0.54**	0.04	**0.50**	0.02
frm	**2.06**	0.02	**1.12**	0.03	**0.84**	0.04	**0.70**	0.01
state	**8.20**	0.23	**5.88**	0.19	**4.28**	0.12	**3.82**	0.15
phn	**18.44**	0.36	**13.46**	0.32	**10.24**	0.29	**6.90**	0.25
word	**30.74**	0.26	**25.00**	0.16	**20.34**	0.18	**17.18**	0.09
orig.	**39.00**		**34.90**		**32.10**		**30.90**	

Far mic

	1g		2g		4g		8g	
	μ	SE	μ	SE	μ	SE	μ	SE
sim	**2.10**	0.02	**0.90**	0.03	**0.60**	0.02	**0.50**	0.01
frm	**7.10**	0.07	**3.40**	0.07	**2.15**	0.18	**1.50**	0.01
state	**27.80**	0.07	**23.55**	0.25	**18.30**	0.01	**15.35**	0.18
phn	**48.80**	0.28	**42.50**	0.21	**34.80**	0.85	**29.20**	0.07
word	**60.80**	0.14	**55.10**	0.35	**49.50**	0.07	**43.75**	0.18
orig.	**67.50**		**65.10**		**62.10**		**61.60**	

Mismatched conditions

	1g		2g		4g		8g	
	μ	SE	μ	SE	μ	SE	μ	SE
sim	**69.60**	0.34	**61.50**	0.32	**41.95**	0.27	**34.85**	0.28
frm	**59.50**	0.32	**46.75**	0.25	**31.90**	0.21	**24.75**	0.39
state	**72.90**	0.35	**68.10**	0.25	**54.10**	0.28	**49.40**	0.32
phn	**78.40**	0.37	**74.90**	0.27	**65.80**	0.37	**61.60**	0.36
word	**78.85**	0.41	**77.35**	0.33	**71.95**	0.34	**70.00**	0.40
orig.	**81.90**		**80.70**		**78.10**		**77.00**	

8.2 MLP transformed results

Tables 6-8 show the effects of MLP feature transformation on word error rates for the all of the sampling conditions (as well as simulation and the original data). Table 6 gives results for near-field data and near-field models; Table 7 gives results for far-field data and far-field models; and Table 8 gives results for far-field data and near-field models.

Table 6. For near-field data and near-field models, the table shows the effect of transforming MFCCs with a phonetically and discriminantly trained MLP. Nine acoustic frames are used as input for the MLP. The "+" symbol indicated augmentation of the MFCC (including 1st and 2nd order deltas) with the MLP features. The models use a single Gaussian per triphone state.

Feature	MFCC	MFCC-MLP	Rel. Imp. to MFCC	MFCC + MFCC-MLP	Rel. Imp. to MFCC
sim	1.4	0.6	57.1%	0.5	64.3%
frame	1.9	1.02	46.3%	0.78	58.9%
state	9.6	6.1	36.5%	5.4	43.8%
phone	21.4	16.1	24.8%	14.6	31.8%
word	37.6	34.4	8.5%	31.3	16.8%
original	44.7	42.3	5.3%	39.6	11.4%

Table 7. For far-field data and far-field models, the table shows the effect of transforming MFCCs with a phonetically and discriminantly trained MLP.

Feature	MFCC	MFCC-MLP	Rel. Imp. to MFCC	MFCC + MFCC-MLP	Rel. Imp. to MFCC
sim	1.8	0.73	59.4%	0.5	72.2%
frame	3.4	2.8	17.6%	1.15	66.1%
state	23.2	19.5	15.9%	15.0	35.3%
phone	45.5	38.7	14.9%	35.6	21.7%
word	63.5	60.4	4.9%	57.3	9.7%
original	71.4	72.2	-1.1%	67.3	5.7%

Table 8. For far-field data and near-field models, the table shows the effect of transforming MFCCs with a phonetically and discriminantly trained MLP. Presumably the use of multiple frames for the MLP reintroduces statistical dependence, and the discriminant MLP training may also increase the fitting to the training set, which differs from the test set.

Feature	MFCC	MFCC-MLP	Rel. Imp. to MFCC	MFCC + MFCC-MLP	Rel. Imp. to MFCC
sim	13.5	71.2	-427.4%	22.3	-65.1%
frame	23.9	71.4	-198.7%	35.5	-48.5%
State	44.2	78.7	-78%	49.2	-11.3%
Phone	58.6	80.4	-37.2%	58.8	-0.3%
word	68.4	80.5	-17.7%	68.6	-0.3%
original	72.2	82.5	-14.2%	73.0	-1.1%

9.0 APPENDIX B – DEMOGRAPHIC INFORMATION FOR SURVEY

The makeup of our participants was self-selected by the snowball process. As shown in Figure 12, more than 50 of our participants were from industry, while slightly fewer classified themselves as being associated with academia. More than ten interviewees identified themselves as working for government. The numbers add up to more than the 86 interviewees, as some had more than one role. At the end of each survey, when we asked the interviewees to give us the names of two additional people who might participate in the survey, we made it clear that they didn't have to limit the type of person they were recommending. Therefore, we believe that the makeup of our survey approximates the makeup of people working in the speech and language technology area.

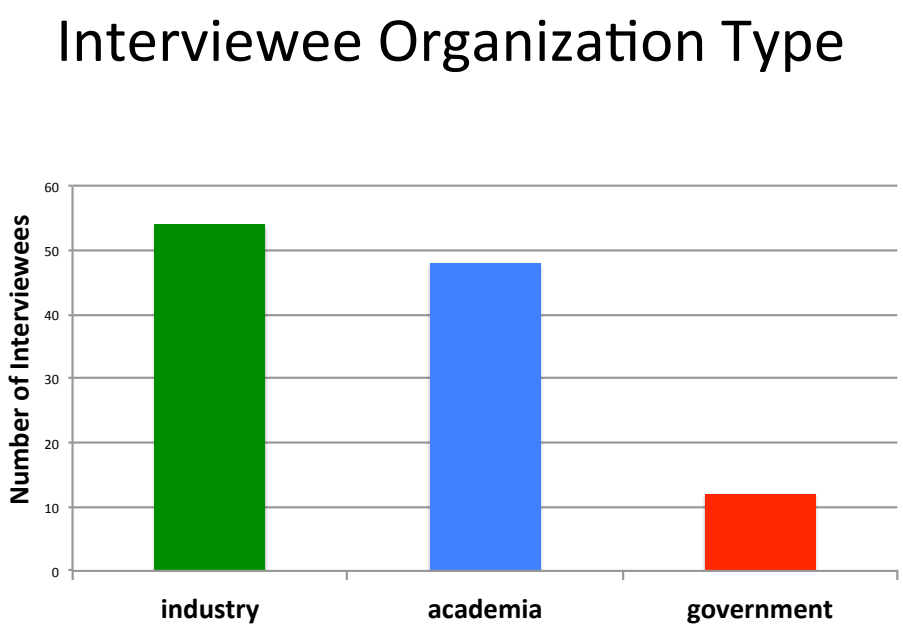

Figure 12. Distribution of interviewees by organization type

The ages of our interviewees (Figure 13) were evenly represented between 40 and 65 years of age. Interviewees tended to recommend people with a substantial background in the research field, and this accounts for the dearth of younger participants. In fact, this was a recurring theme in our interviews, as the general perception was that there were not a lot of younger entrants into the field, and that this was an issue that needed to be addressed.

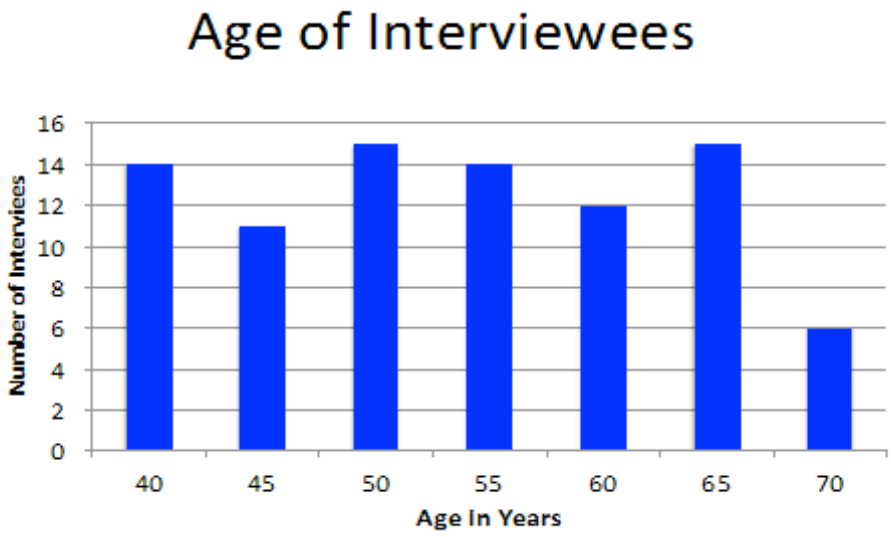

Figure 13. Distribution of interviewees by age. Ages were rounded to the nearest 5 years, so "40" represents ages from 37.5 to 42.5 years old. "70" refers to those aged 67.5 and above.

We then asked our interviewees what their current job, or "professional affiliation" was. Figure 14 shows the job categories as self-reported.

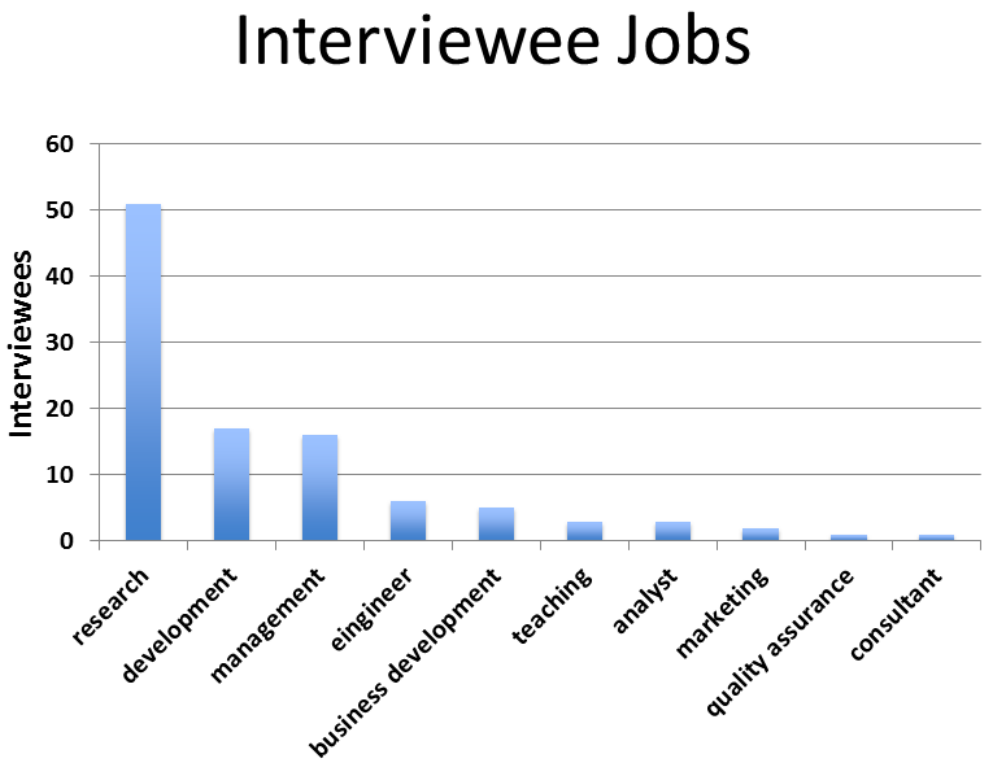

Figure 14. Distribution of interviewees by job type; subjects sometimes identified themselves as working in more than one area (e.g., research and teaching).

Our interviewees identified more than 12 technology areas in which they are currently working in, although there is substantial overlap in the categories. This was particularly true if the interviewee was working at a speech technology company or in a speech research group within industry. Additionally, there were a smattering of management personnel, and a few analysts and consultants, whose work in the field is more varied than those doing research or development. The resulting distribution is shown in figure 15.

The predominant identification was work in automatic speech recognition. However categories of mobile-embedded, acoustics, keyword spotting, and language modeling could also have been considered ASR. The other categories included text-to-speech, human-computer interfaces, and various identification tasks (language, speaker, and other biometrics).

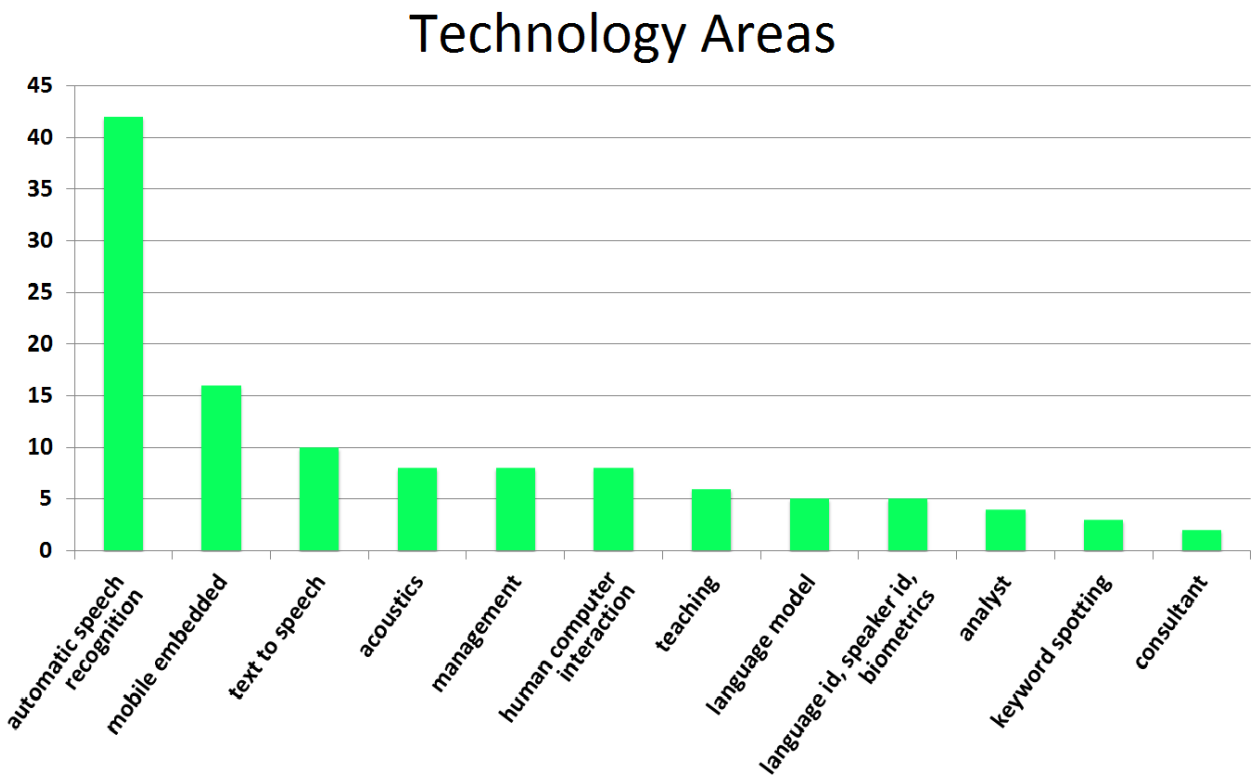

Figure 15. Distribution of interviewees by current work area

10.0 APPENDIX C - BIBLIOGRAPHY – OTHER RELEVANT PUBLICATIONS

The reviews given in the main body of the report have extensive bibliographies of papers and books in the speech and language technology arena. These are readily available, and we have not attempted to replicate or mimic them here.

In the process of our surveys we located many other papers that seemed relevant to our study, and in particular the experts interviewed in our community survey also recommended a number

of other significant papers. Here are a few of the more instructive additional documents that we found.

a. An early paper, [40], is interesting because of the breadth of adaptation methods known at that time, and which are still used in our modern systems. The paper outlines MAP estimation, Cepstral Mean Normalization, MLLR, Vector Field Smoothing, VTLN, and Speaker Adaptive Training. It is stunning that these basic methods are still in use, and that they have been evolved over the past two decades, without fixing the basic performance issues in ASR.

b. An intriguing exercise in modern speech recognition system building may be found in [41]. The author carefully exercises modern acoustic and language models, including language model and acoustic adaptation, to build a system for one talker. The descriptions are succinct and clear. The performance of the resulting system is typical: 20 to 40% word error for simple sentences. This is a detailed and dismal view of current technology.

c. There have been attempts to field the current ASR technology in large scale applications. [42] describes the monumental effort to create speech recognition for voice search at Google, along with the appropriate user interface and other infrastructure. The bottom line is that this indefatigable creator of technology has created a speech recognizer with a word error rate of 17%. While this performance is apparently commercially viable, it gives the technician looking for success substantial heartache. It means that the current technology, fed by essentially infinite data and compute, is substantially defective!

d. There is a move afoot to look at robustness directly. In [43], the authors show that large margin measures create slightly more robust processes. It is encouraging to see movement towards robustness, but broader statistics, while part of the answer, ignores the difficulties in the models themselves.

e. An attempt to move away from the simple HMM models dating from the 1960's may be seen in [44]. The authors attempt to use an underlying hidden generative model and demonstrate improved performance on phonetic recognition of TIMIT. This movement away from simple HMM states points out the potential gain from more sensible models of the speech generation process.

f. Adaptation using unsupervised data is described in [45], creating a "robust" system on-the-fly. They demonstrate that transcription is not a necessary part of adaptation.

We asked our informants about papers or books which might inform a reader about the issues in the state of the art performance, or which were particularly enlightening about the engineering or scientific issues. Most declined to make a recommendation, but several interesting suggestions were made. We list them here:

1. The recent writings and papers of Larry Gillick
    a. D. Gillick and S. Wegmann, L. Gillick. Discriminative Training for Speech Recognition is Compensating for Statistical Dependence on the HMM Framework. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), pp. 4745-4748, Kyoto, Japan, March 2012.
    b. D. Gillick, L. Gillick, and S. Wegmann. Don't Multiply Lightly: Quantifying Problems with the Acoustic Model Assumptions in Speech Recognition. Proceedings of the

Automatic Speech Recognition and Understanding Workshop (ASRU 2011), pp. 71-76, Big Island, Hawaii, December 2011.
   c. S. Wegmann and L. Gillick. Why Has (Reasonably Accurate) Automatic Speech Recognition Been So Hard to Achieve? ArXiv.org CoRR abs/1003.0206, February 2010.
2. Good's paper on smoothing in Biometrica
   a. I. J. Good. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, Vol. 40, No. 3-4, pp. 237-264, December 1953.
3. The papers of Dan Povey
   a. O. Vinyals and D. Povey. Krylov Subspace Descent for Deep Learning. Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012), La Palma, Canary Islands, April 2012.
   b. D. Povey, M. Hannemann et al . Generating exact lattices in the WFST framework. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan, March 2012.
   c. K. Reidhammer, T. Bocklet, A. Ghoshal, and D. Povey. Revisiting Semi-continuous Hidden Markov Models. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan, March 2012.
   d. N. T.Vu, T. Schultz, and D. Povey. Modeling Gender Dependency in the Subspace GMM Framework. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan, March 2012.
   e. O. Vinyals, S.V. Ravuri, and D. Povey. Revisiting Recurrent Neural Networks for Robust ASR. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan, March 2012.
   f. D. Povey, A. Ghoshal, et al. The Kaldi Speech Recognition Toolkit. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2011), Big Island, Hawaii, December 2011.
   g. D. Povey, G. Zweig, and A. Acero. Speaker Adaptation with an Exponential Transform. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2011), Big Island, Hawaii, December 2011.
   h. D. Povey, L. Burget, et al. The Subspace Gaussian Mixture Model– a Structured Model for Speech Recognition. *Computer Speech and Language*, Vol. 25, Issue 2, pp. 404-439, April 2011.
   i. D. Povey and K. Yao. A basis representation of constrained MLLR transforms for robust adaptation. *Computer Speech and Language*, Vol. 26, Issue 1, pp. 35-51, January 2012.
   j. H. Xu, D. Povey, L. Mangu and J. Zhu. Minimum Bayes Risk decoding and system combination based on a recursion for edit distance. *Computer Speech and Language*, Vol. 25, Issue 4, pp. 802-828, October 2011.
   k. D. Povey and K. Yao. A Basis Method for Robust Estimation of Constrained MLLR. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), Prague, Czech Republic, May 2011.
   l. D. Povey, M. Karafiat, A. Ghoshal, and P. Schwarz. A Symmetrization of the Subspace Gaussian Mixture Model. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), Prague, Czech Republic, May 2011.
   m. Y. Qian, D. Povey and J. Lu. State-Level Data Borrowing for Low-Resource Speech Recognition Based on Subspace GMMs. Proceedings of the 12th Annual Conference of

the International Speech Communication Association (Interspeech 2011), Florence, Italy, August 2011.
4. Colin Cox – Statistical Significance
   a. D.R. Cox. Statistical Significance Tests. *British Journal of Clinical Pharmacology*, Vol. 14, pp. 325-331, 1982.
5. HMMs for Speech Recognition by Huang et al
   a. X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, 1990.
6. Kai Fu Li Thesis
   a. K. F. Li. *The Development of the SPHINX Recognition System*. Springer, October 1988.
7. Holmes and Mattingly writings (a small selection noted here)
   a. J. Holmes, I.G. Mattingly, and J.N. Shearme. Speech synthesis by rule. *Language and Speech*, Vol. **7**, No. 3, pp.127-143, 1964.
   b. I.G. Mattingly. Synthesis by rule as a tool for phonological research. *Language and Speech*, Vol. 14, No. 1, pp. 47-56, 1971.
   c. J. Holmes. Formant synthesizers, cascade or parallel. *Speech Communications*, Vol. 2, pp. 251-273, 1983.
   d. J. Holmes. Influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE Transactions on Audio Electroacoustics*, Vol. 21, Issue 3, pp. 298-305, June 1973.
8. David MacKay on Information Theory and Algorithms
   a. D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, various years.
9. Bourlard, Hermansky, and Morgan on daring to risk increasing the error rate by trying radically new ideas
   a. H. Bourlard, H. Hermansky, and N. Morgan. Towards increasing speech recognition error rates. *Speech Communication*, Vol. 18, pp. 205–231, 1996.
10. Bridle and Richards on Hidden Dynamic Models
    a. H. B. Richards and J. S. Bridle. The HDM: A Segmental Hidden Dynamic Model of Coarticulation. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999), Phoenix, Arizona, 1999.
11. Deep Neural Network papers from Microsoft
    a. G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, Issue 1, pp. 30-42, January 2012.
12. Li Deng's writings (These are extensive – this is an important example, but there are literally hundreds of references)
    a. L. Deng, Dynamic Speech Models—Theory, Algorithm, and Application (book review). *IEEE Transactions on Neural Networks*, Vol. 20, Issue 3, March 2009.
13. Miami Children's Hospital
    a. A. T. Winfree. When time breaks down – the story of fractals. *When Time Breaks Down: The Three-Dimensional Dynamics of Electrochemical Waves and Cardiac Arrhythmias*, Princeton University Press, April 1987.
14. The autobiography of Craig Venter
    a. J. C. Venter. *A Life Decoded: My Genome: My Life*. Penguin, 2007.

11.0 LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ASR: Automatic Speech Recognition

DNN: Deep Neural Network

DTW: Dynamic Time Warp

ECoG: Electrocorticography

FF: Far-field

GMM: Gaussian Mixture Model

HMM: Hidden Markov Model

HTK: HMM Tool Kit

IVR: Interactive Voice Response

LM: Language Model

LVCSR: Large Vocabulary Continuous Speech Recognition

MFCC: Mel Frequency Cepstral Coefficient

ML: Maximum Likelihood

MLLR: Maximum Likelihood Linear Regression

MLP: Multi Layer Perceptron

MPE: Minimum Phone Error

NF: Near-field

NIST: National Institute of Standards and Technology

RASTA: RelAtive SpecTral Analysis

ROVER: Recognizer Output Voting Error Reduction

STT: Speech To Text

VTLN: Vocal Tract Length Normalization

VUI: Voice User Interface

WER: Word Error Rate