COLLECTION AND ANALYSIS OF DATA FROM REAL USERS: IMPLICATIONS FOR SPEECH RECOGNITION/UNDERSTANDING SYSTEMS

Judith Spitz and the Artificial Intelligence Speech Technology Group

NYNEX Science and Technology, Research and Development
500 Westchester Avenue
White Plains, New York 10604

ABSTRACT

Performance estimates given for speech recognition/understanding systems are typically based on the assumption that users will behave in ways similar to the observed behavior of laboratory volunteers. This includes the acoustic/phonetic characteristics of the speech they produce as well as their willingness and ability to constrain their input to the device according to instructions. Since speech recognition devices often do not perform as well in the field as they do in the laboratory, analyses of real user behavior have been undertaken. The results of several field trials suggest that real user compliance with instructions is dramatically affected by the particular details of the prompts supplied to the user. A significant amount of real user speech data has been collected during these trials (34,000 utterances, 29 hours of data). These speech databases are described along with the results of an experiment comparing the performance of a speech recognition system on real user vs. laboratory speech.

INTRODUCTION

Speech recognition/understanding systems will ultimately establish their usefulness by working well under real application conditions. Success in the field will depend not only on the technology itself but also on the behavior of real users. Real user behavior can be characterized in terms of 1. what people say and 2. how they say it.

What people say: Real user compliance

Until the advent of a high performance continuous speech, unconstrained vocabulary/grammar, interactive speech understanding system, users must constrain their spoken interactions with speech recognition/understanding systems. Constraints may require speaking words in isolation, conforming to a limited vocabulary or grammar, restricting queries to a particular knowledge domain, etc. The users' willingness to comply with instructions specifying these constraints will determine the success of the technology. If users are willing or even able to confine themselves to one of two words (e.g., yes or no), a two-word speech recognition system may succeed. If users are non-compliant (e.g., say the target words embedded in phrases, say synonyms of the target words, reject the service as a result of the constraining instructions), the technology will fail in the field; despite high accuracy laboratory performance.

How compliant are real users? The answer may be application-specific, dependent on particulars such as 1. frequency of repeat usage of the system, 2. motivation of the users, 3. cost of an error, 4. nature of the constraint, etc. It would be useful to understand the factors that predict compliance, and to know whether generalizations can be made across applications. In addition, it would be useful to have a better understanding of how to maximize user compliance.

Moreover, there is value in analyzing non-compliant behavior. To the extent that non-compliance takes the form of choosing synonyms of the target words, the recognizer's vocabulary must be expanded. If non-compliance takes the form of embedding the target word in a phrase, word spotting or continuous speech recognition is required. If non-compliance is manifested by the user consistently wandering outside the knowledge domain of the speech recognition/understanding system, better instructions may be required. Data from real users should provide researchers and developers with the information necessary to both specify and develop the technology required for successful deployment of speech recognition/understanding systems.

How people speak: Real user speech

It seems intuitively obvious that to maximize the probability of successfully automating an application with speech recognition, a recognizer should be trained and tested on real user speech. This requires the collection of data from casual users interacting with an automated or pseudo-automated system, thereby producing spontaneous goal-directed speech under application conditions. These databases can be difficult and expensive to collect and so it is not surprising that speech recognition systems are most typically trained and tested on speech data collected under laboratory conditions. Laboratory databases can be gathered relatively quickly and inexpensively by recording speech produced by cooperative volunteers who are aware that they are participating in a data collection exercise. But these databases typically have relatively few talkers and speech that is recited rather than spontaneously-produced.

Potential differences between real user and laboratory speech databases would be of little interest if speech recognition systems were performing as well in field applications as they are in the laboratory. However, there is data to suggest that this is not the case; systems performing well in the laboratory often achieve significantly poorer results when confronted with real user data [1,2].

A number of features that differentiate real user from laboratory database collection procedures may have an impact on the performance of speech recognition systems. One that has received specific attention in the literature is that of spontaneously-produced vs. read speech. Jelinek et al. [3] compared the performance of a speech recognition system when tested on pre-recorded, read and spontaneous speech produced by five talkers. Results indicate decreasing performance for the three sets of test material (98.0%, 96.9% and 94.3% correct, respectively). Rudnicky et al. [4] evaluated their speech recognition system on both read and spontaneous speech produced by four talkers and found that performance was roughly equal for the two data sets (94.0% vs. 94.9% correct, respectively). It is important to note, however, that the spontaneous speech used for this comparison was "live clean speech" defined as "only those utterances that both contain no interjected material (e.g., audible non-speech) and that are grammatical". Degradation in performance was indeed seen when the test set included all of the "live speech" (92.7%). Zue et al. [5] also evaluated their speech recognition system on read and spontaneous speech samples. Word and sentence accuracy were similar for the two data sets. For each of these studies, 'real user' speech samples were recorded under wideband application-like conditions. For at least two of the studies ([4], [5]), the 'real users' were apparently aware that they were participating in an experiment.

It has not been possible to collect databases that are matched with respect to speakers for telephone speech, probably because the anonymity of the users of telephone services makes it difficult to obtain read versions of spontaneously-produced speech from the same set of talkers. Therefore, there is little published data on the effects of read vs. spontaneous speech on the performance of recognition systems for telephone applications. Differences in speakers not withstanding, there is recent data to suggest that recognition performance can be significantly poorer when testing on real user telephone speech as compared to tests using telephone speech collected under laboratory conditions ([1], [2]).

In summary, laboratory and real user behavior can be characterized along at least two important dimensions: compliance and speech characteristics. To gain a better understanding of how to improve the field performance of speech recognition/understanding systems, we have been collecting and analyzing both laboratory and real user data. The goal of this paper is to summarize our work in the analysis of 1. real user compliance for telephone applications and 2. laboratory vs. real user speech data for the development of speech recognition/understanding systems.

REAL USER DATABASE COLLECTION PROCEDURES

Three real user telephone speech databases have been collected by pseudo-automating telephone operator functions and digitally recording the speech produced by users as they interacted with the services. In each case, experimental equipment was attached to a traditional telephone operator workstation and was capable of : 1. automatically detecting the presence of a call, 2. playing one of a set of pre-recorded prompts to the user, 3. recording user speech, 4. automatically detecting a user hang-up and 5. storing data about call conditions associated with a given speech file (e.g., time of day, prompt condition, etc.). The three operator services under study were 1. Intercept Services (IS) 2. Directory Assistance Call Completion (DACC) and 3. Directory Assistance (DA). In addition to collecting data for several automated dialogues, recordings were made of traditional 'operator-handled' calls for the services under investigation.

Each of these databases was collected in a real service-providing environment. That is, users were unaware that they were participating in an experiment. The identity of the speakers was not known, so a precise description of dialectal distribution is difficult. Calls reached the trial position through random assignment of calls to operator positions, a task performed by a network component known as an Automatic Call Distributor (ACD). Therefore, for each of the databases, it is assumed that the number of utterances corresponds to the number of speakers. We have so far collected nearly 29 hours of real user speech: 34,000 utterances (presumably from that many different speakers).

REAL USER COMPLIANCE

For the IS trial, users were asked what telephone number they had just dialed. For the DACC trial, users were asked to accept or reject the call completion service. For the DA trial, users were asked for the city name corresponding to their directory request. The 'target' responses, therefore, were digit strings, yes/no responses and isolated city names, respectively. Users were presented with different automated prompts varying along a number of dimensions. Their responses were analyzed to determine the effects of dialogue condition on real user compliance (frequency with which users provided the target response).

Intercept Services: 'Simple' Digit Recognition

One problem with digit recognition is that users may say more than just digits. The target response for the IS trial was a digit string. The automated prompts to the users varied with respect to the 1. presence/absence of an introductory greeting which informed users that they were interacting with an automated system 2. speed with which the prompts were spoken (fast, slow), and 3. the explicitness of the prompts (wordy, concise). In addition, data were recorded under an operator-handled condition. During operator-handled intercept calls, operators ask users, "What number did you dial?".

A total of 3794 utterances were recorded: 2223 were in the automated-prompt conditions and 1571 were in the operator-handled condition. 'Non-target' words were defined as anything other than the digits '0' through '9' and the word 'oh'. Results showed that only 13.6% of the utterances in the automated conditions were classified as non-target, while 40.6% of the utterances in the operator-handled condition fell into the non-target category.

Non-target utterances were further classified as '100-type' utterances (that is, utterances in which the user said the digit string as "992-4-one-hundred, etc.) and 'extra verbiage' utterances (that is, utterances in which the user said more than just the digit string such as, "I think the number is ...", or "oh, um, 992 ..."). For both automated and operator-handled calls, users produce more extra verbiage utterances than 100-type utterances. Both types of non-target responses occurred more than twice as often in the operator-handled condition compared to the automated conditions.

The speed and wordiness of the automated prompts did not affect user compliance. However, contrary to our expectations, the data suggest that the proportion of non-target responses is substantially reduced when the user is *not* given an introductory greeting which explains the automated service (19.2% vs. 4.9% non-target responses for the greeting vs. no-greeting conditions, respectively). Instead, giving users an immediate directive to say the dialed number results in the highest proportion of responses which are restricted to the desired vocabulary. It appears that even untrained users are immediately attuned to the fact that they are interacting with an automated service and modify their instinctual response in ways beneficial to speech recognition automation. At least for this application, brevity is best. For more information on this trial, see [6].

Directory Assistance Call Completion: Yes/No Recognition

The target response for the DACC trial was an isolated 'yes' or 'no' response. Successful recognition of these words would have many applications, but the problem for a two-word recognizer is that users sometimes say more than the desired two words. Data were collected under three automated prompt conditions and one operator-handled condition. The operator asked, "Would you like us to automatically dial that call for an additional charge of __ cents?" The three automated prompts were as follows: 1. a recorded version of the operator prompt 2. a prompt which explicitly asked for a 'yes' or 'no' response and 3. a prompt that asked for a 'yes' or hang up response.

A total of 3394 responses were recorded; 1781 were operator-handled calls, while 1613 were calls handled by automated prompts. Figure 1 shows the percentage of 'yes' responses among the affirmative responses as a function of dialogue condition. Results again indicate that variations in the prompt can have a sizable effect on user compliance and that there are considerable differences between user behavior with a human operator vs. an automated system.

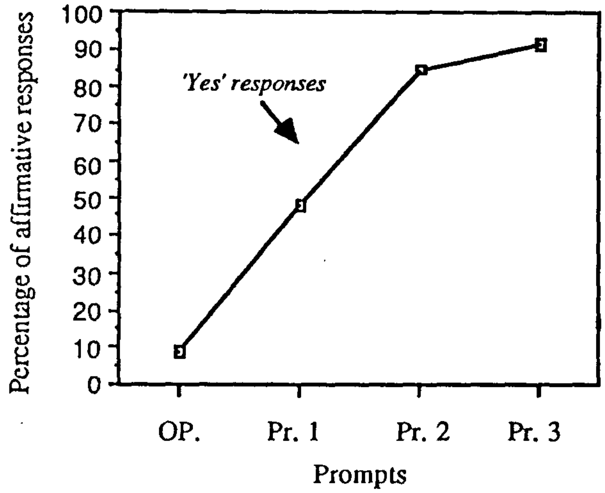

Figure 1: Percentage of affirmative responses that were the target response ('yes') as a function of dialogue condition.

Non-target affirmative responses were categorized as 'yes, please', 'sure' and 'other'. A response was categorized as 'other' if it accounted for less than 5% of the data for any prompt condition. The frequency of occurrence of these non-target responses as a function of dialogue condition is shown in Table 1.

	Operator handled	Prompt 1	Prompt 2	Prompt 3
'yes, please'	22	16	3	1
'sure'	16	10	0	0
other	53	26	12	7

Table 1: Percentage of users' affirmative responses as a function of prompt condition.

The operator-handled condition exhibited the greatest range of variability, with 53% of the affirmative responses falling into the 'other' category. For more information on the DACC trial, see [7].

"Directory assistance, what city please?"

The target response for the Directory Assistance trial was an isolated city name. Data were collected under four automated prompt conditions and one operator-handled condition. Directory Assistance operators typically ask users "What city, please?". One automated prompt used the same wording as the operator; the other three were worded to encourage users to say an isolated city name. Recording was initiated automatically at the offset of a beep tone that prompted users to respond. Recording was terminated by a human observer who determined that the user had finished responding to the automated request for information.

A total of 26,946 utterances were collected under automated conditions. Operator-handled calls were collected during a separate trial [8] and only 100 of these utterances were available for analysis. Figure 2 shows the percentage of target responses as a function of dialogue condition.

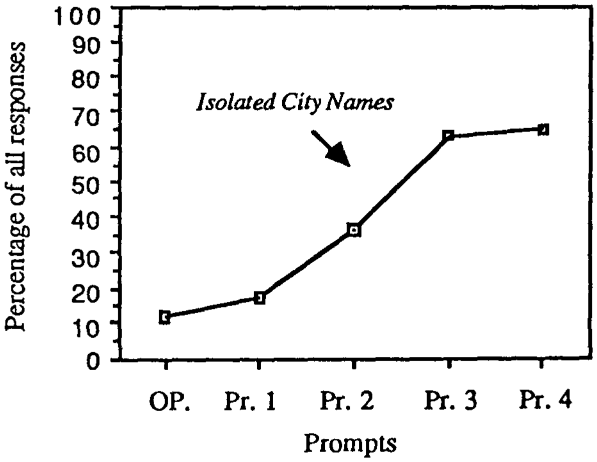

Figure 2: Percentage of all responses that were isolated city names as a function of dialogue condition.

As in the other two trials, user behavior was quite different for operator-handled vs. automated calls. On average, users were almost four times more likely to say an isolated city name in response to an automated prompt than to an operator query. Moreover, the wording of the automated prompt had a large effect on user compliance. Superficially minor variations in prompt wording increased user compliance by a factor of four (15.0% vs. 64.0% compliance for prompt 1 vs. 4, respectively).

Very few users either did not reply or replied without a city name in response to an operator prompt. For the automated conditions, between 14% and 23% of the users simply did not respond. Between 3% and 23% responded without including a city name. To interpret these results, we point out that in contrast to the users of IS and DACC services, Directory Assistance users tend to be repeat callers with well-rehearsed scripts in mind. When the familiar interaction is unexpectedly disrupted, some of these users appear to be unsure of how to respond.

Of particular interest was the effect of dialogue condition on the frequency of occurrence of city names embedded in longer utterances. These results appear in Figure 3.

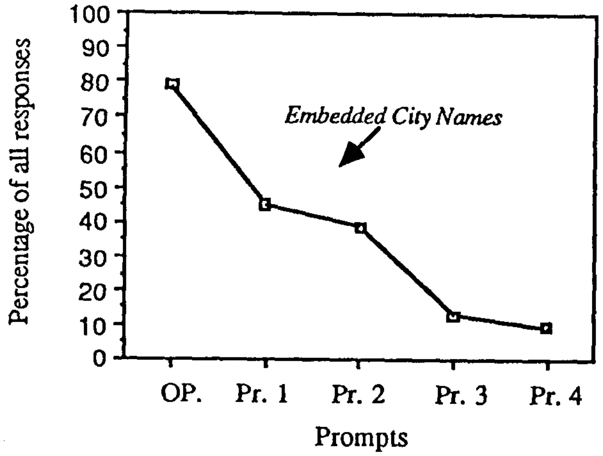

Figure 3: Percentage of all responses that were embedded city names as a function of dialogue condition.

It is clear that embedded responses are most typical during user-operator interactions. To allow for this response mode, a recognizer would have to be able to 'find' the city name in such utterances. This could be accomplished with a word spotting system or with a continuous speech recognition/understanding system. To consider the difficulty of the former, embedded city name responses were further categorized as simple vs. complex; assuming that the former would be relatively easy to 'spot'. A 'simple' embedded city name was operationally defined as a city name surrounded by approximately one word (for example, "Boston, please", "um, Boston", "Boston, thank you"). The proportion of embedded utterances classified as 'simple' as a function of prompt is shown in Figure 4.

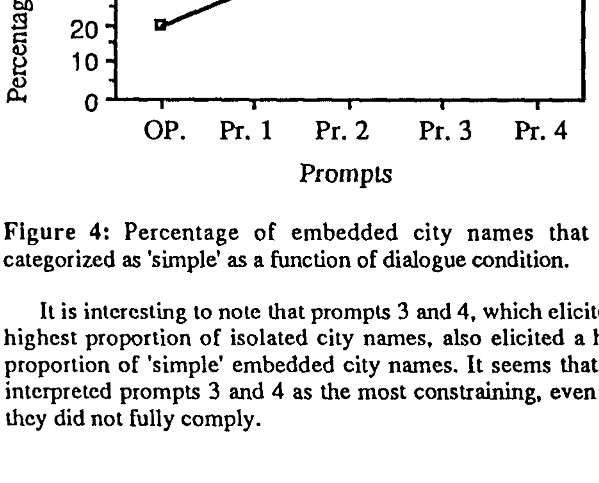

Figure 4: Percentage of embedded city names that were categorized as 'simple' as a function of dialogue condition.

It is interesting to note that prompts 3 and 4, which elicited the highest proportion of isolated city names, also elicited a higher proportion of 'simple' embedded city names. It seems that users interpreted prompts 3 and 4 as the most constraining, even when they did not fully comply.

DISCUSSION

The results of this series of experiments on real user compliance suggest that this aspect of user behavior is significantly different when interacting with a live operator than when interacting with an automated system. The lesson is that feasibility projections made on the basis of observing operator-handled transactions will significantly underestimate automation potential. In addition, the precise wording of the prompts used in a speech recognition/understanding application significantly affects user compliance and therefore the likelihood of recognition success. Users seem to know immediately that they are interacting with an automated service and explicitly informing them of this fact does not improve (in fact, decreases) user compliance. Prompts should be brief and the tasks should not be too unnatural. Although not discussed above, informal analysis of the data suggests that very few users attempted to interrupt the prompts with their verbal responses. While this would suggest that 'barge-in' technology is not a high priority, it should be noted that the users under investigation were all first-time users of the automated service. It seems likely that their desire to interrupt the prompt will increase with experience, as has been found for Touch-Tone applications.

Although each of the applications under investigation was different with respect to the degree of repeat usage, the motivation of the user, the cost of an error, etc., the trials were similar in that there was no opportunity for learning on the part of the user. This is an important factor in the success of many speech recognition/understanding systems and is an area of future research for the group.

LABORATORY DATABASE COLLECTION PROCEDURES

While real user speech databases provide value to the researcher/developer, they present limitations as well. Most notably, the a priori probabilities for the vocabulary items under investigation are typically quite skewed. It is rare, in a real application, that any one user response is as likely as any other. The DA data collection gathered almost 27,000 utterances, yet there are less than 10 instances of particular cities and, correspondingly, less than 10 exemplars of certain phones. If these data are to be used for training speech recognition/understanding systems, they must be supplemented with laboratory data. To this end, as well as for the purposes of comparing real user to laboratory data, application-specific and standardized laboratory telephone speech data collection efforts were undertaken.

Application-specific laboratory speech database collection

A laboratory city name database was collected by having volunteers call a New York-based laboratory from their New England-based home or office telephones. Talkers were originally from the New England area and so were assumed to be familiar with the pronunciation of the target city names.

When a speaker called, the system asked him/her to speak the city names, waiting for a prompt before saying the next city name (the order of the city names was randomized so as to minimize list effects). 10,900 utterances from over 400 speakers have been collected in this way.

This kind of database provides some of the characteristics of a real user database (a sample of telephone network connections and telephone sets). The speech, however, is read rather than spontaneously-produced and the speakers are aware that they are participating in a data collection exercise. This database has been compared to the DA corpus just described. Results are reported below.

Standardized telephone speech data collection

The TIMIT database reflects the general nature of speech and is not designed for any particular application [10]. It is well known that the telephone network creates both linear and nonlinear distortions of the speech signal during transmission. In the development of a telephone speech recognition/understanding system, it is desirable to have a database with the advantages of the TIMIT database, coupled with the effects introduced by the telephone network. Towards this end, a data collection system has been developed to create a telephone network version of the TIMIT database (as well as other standardized wideband speech databases). The system is capable of 1. systematically controlling the telephone network and 2. retaining the original time-aligned phonetic transcriptions.

Figure 5 shows the hardware configuration used in the collection of the NTIMIT (Network TIMIT) database. The TIMIT utterance is transmitted in an acoustically isolated room through an artificial mouth. A telephone handset is held by a telephone test frame mounting device. Taken together, this equipment is designed to approximate the acoustic coupling between the human mouth and the telephone handset. To allow transmission of utterances to various locations, "loopback" devices in remote central offices were used.

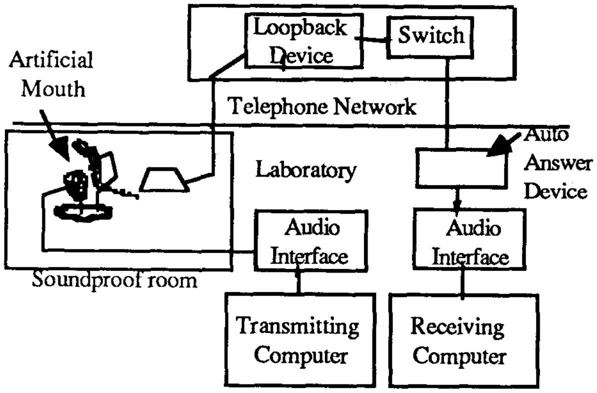

Figure 5: Hardware configuration for NTIMIT database collection.

The choice of where to send the TIMIT utterances was carefully designed to ensure geographic coverage as well as to keep the distribution of speaker gender and dialect for each geographic area roughly equivalent to the distribution in the entire TIMIT database. To obtain information about transmission characteristics such as frequency response, loss, etc., two calibration signals (sweep frequency and 1000 Hz pure tone signals) were sent along with the TIMIT utterances. NTIMIT utterances were automatically aligned with the original TIMIT transcriptions.

The NTIMIT database is currently being used to train a telephone network speech recognition system. Performance will be compared to that of a system trained on a band-limited version of the TIMIT database to determine the effects of a 'real' vs. simulated telephone network on recognition results. In addition, we are evaluating the performance of a recognizer trained on a combination of material from real user speech databases and NTIMIT.

For more information on NTIMIT, see [9]. The NTIMIT database is being prepared for public distribution through NIST.

LABORATORY VS. REAL USER SPEECH FOR TRAINING AND TESTING SPEECH RECOGNITION SYSTEMS

The laboratory and real user city name databases described above allowed us to evaluate the performance of a speaker independent, isolated word, telephone network speech recognition system when tested on laboratory vs. real user data. Two training scenarios were included: 1. trained on laboratory data and 2. trained on real user data. To equate the number of training and testing samples for each of the 15 city names under investigation, only a subset of each database was used (and only isolated city names were used from the real user database).

Each database was divided into a training and testing set, consisting of 90% and 10% of the databases, respectively. A phonetically-based speaker independent isolated word telephone network speech recognition system was used for this experiment. The recognizer, developed as part of an MIT-NYNEX joint development project, was built upon a system developed at MIT (for more details on the MIT recognizer, see [11]). The system was trained on each training set and then tested on each testing set. This resulted in four training/testing conditions.

Results revealed that performance changed little as a function of laboratory vs. user training databases when tested on laboratory speech (95.9% vs. 91.1% for laboratory and user training databases, respectively). In contrast, performance changed dramatically as a function of training database when tested on real user speech (52.0% vs. 87.7% for laboratory and user training databases, respectively). Two points of interest here are: 1. The recognizer that was trained and tested on laboratory speech performed almost 9% better than the recognizer trained and tested on real user speech (95.9% vs. 87.7% respectively). Apparently, recognizing real user speech is an inherently more difficult problem. 2. Performance of the laboratory-trained system on real user speech was 43.9% poorer than the same system tested on laboratory speech. A number of experiments were conducted to better understand these results.

It is assumed that the performance of the real user-trained system on real user speech (87.7%) represents optimal performance. Therefore, the performance discrepancy to be explored is the difference between 52.0% (the lab-trained system on real user speech) and 87.7%. Each one of the recognizer's components involved in the training was considered for analysis. This included 1. phonetic acoustic models, 2. silence acoustic models 3. lexical transition weights and 4. lexical arc deletion weights. A series of experiments were done in which each of these components from the real user-trained recognizer was systematically substituted for its counterpart in the laboratory-trained recognizer. The resulting hybrid recognizer was evaluated at each stage.

Results revealed that an overwhelming majority of the performance difference could be accounted for by the acoustic models for silence. A recognizer trained on laboratory speech which used silence acoustic models trained on real user speech achieved 82% accuracy when tested on real user speech. An acoustic analysis of the two databases revealed that they were quite similar with respect to the frequency characteristics of the non-speech portions of the signal and the signal-to-noise ratios. Rather, it was the mean duration and variability in duration of the non-speech signal prior to the onset of the speech that accounts for this effect. It is important to note that the silence surrounding the laboratory-collected city names was artificially controlled by both the data collection procedures (talkers knew they had only a limited amount of time to speak before hearing the prompt to read the next city name) and subsequent hand editing. The real user data were not since a field recognizer will not see controlled or hand edited data. While these results may appear to be artifactual, they point out the limitations imposed on the researcher/developer in only being exposed to laboratory data. Further experimentation revealed that using real user-trained phonetic acoustic models accounts for most of the remaining 6%, with decreasing importance attributable to real user-trained lexical transition weights and real user-trained lexical arc deletion weights.

DISCUSSION AND CONCLUSIONS

The longer term goal of the work summarized above is to develop speech recognition/understanding systems that maintain high accuracy performance in real telephone applications.

Comparisons between a real user and a laboratory speech database highlight how apparently superficial differences between the two database types can result in dramatic differences in recognition performance. Having accounted for this kind of effect, there appears to be an approximately 6% performance difference that can be attributed to differences in the speech signal itself, even for a small vocabulary isolated-word recognition task. This is the subject of further investigation. The differences reported in the literature when comparing laboratory to field performance for continuous speech recognition typically exceeds 6% ([1], [2]). It seems likely that differences between read and spontaneous speech are minimized in the production of isolated words. Future research will include continued study of the differences between real user and laboratory speech and its effects on recognition/understanding performance for a continuous speech task.

ACKNOWLEDGMENTS

I would like to gratefully acknowledge the members of the AI Speech Group: Sara Basson, Ben Chigier, Charles Jankowski, Ashok Kalyanswamy, Suzi Levas, David Lubensky, Kenneth Nawyn, John Pitrelli, Lisa Russell, Kim Silverman and Dina Yashchin for their participation in the projects described and for their editorial assistance in writing this paper; the MIT Spoken Language Systems Group who developed the base recognition system and assisted in various ways towards the completion of this work; the NYNEX Systems Analysis Laboratory, Maryann Cavan, John Mitchell, Erik Urdang and the employees of New England Telephone Operator Services who made the collection of these databases possible.

REFERENCES

[1] Yashchin, D., Basson, S., Lauritzen, N., Levas, S., Loring, A., Rubin-Spitz, J. (1989) Performance of Speech Recognition Devices: Evaluating Speech Produced Over the Telephone Network, Proceedings of ICASSP, Vol. 1, S10b.10, p. 552-555, May 1989.

[2] Bounds, A., Prusak, M. (1989) Implementing Speech Recognition in an Operational Environment, Proceedings of AVIOS 1989, p. 3-7 and viewgraph presentation.

[3] Jelinek, F., Speech Recognition Group (1985) A Real-Time, Isolated-Word, Speech Recognition System for Dictation Transcription, Proceedings of ICASSP, Vol. 2, 23.5.1, p. 858-861, March, 1985.

[4] Rudnicky, A.I., Sakamoto, M. and Polifroni, J. H. (1990) Spoken Language Interaction in a Goal-Directed Task, Proceedings of ICASSP, Vol. 1, S2.2, p. 45-48, April, 1990.

[5] Zue, V., Daly, N., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., Seneff, S. and Soclof, M. (1989) Preliminary Evaluation of the Voyager Spoken Language System, Proceedings of the Second DARPA Speech and Natural Language Workshop, October, 1989.

[6] Rubin-Spitz, J., Yashchin, D. (1989) Effects of Dialogue Design on Customer Responses in Automated Operator Services , Proceedings of Speech Tech, 1989.

[7] Basson, S., Christie, O., Levas, S., Spitz, J. (1989) Evaluating Speech Recognition Potential in Automating Directory Assistance Call Completion, 1989 AVIOS Proceedings.

[8] Altom, M.J., Velius, G. (1990) Estimating Directory Assistance Savings with Automatic City Name Recognition, Bellcore Technical Memorandum TM-ARH015286.

[9] Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database (1990), Proceedings of ICASSP, Vol. 1, S2.19. p. 109-112, April, 1990.

[10] Fisher, W., Doddington, G.R., Goudie-Marshall, M (1986) The DARPA Speech Recognition Research Database: Specifications and Status, Proceedings of the DARPA Workshop on Speech Recognition, February, 1986.

[11] Zue, V., Glass, J., Phillips, M., Seneff, S. (1989) The MIT SUMMIT Speech Recognition System: A Progress Report, Proceedings of the First DARPA Speech and Natural Language Workshop, p. 178-189.