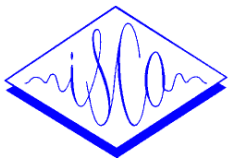

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

ARE LABORATORY DATABASES APPROPRIATE FOR TRAINING AND TESTING TELEPHONE SPEECH RECOGNIZERS?

Benjamin Chigier and Judith Spitz

NYNEX Artificial Intelligence Speech Technology Group
500 Westchester Avenue, White Plains, NY 10604

ABSTRACT

Automatic speech recognition systems are typically trained on speech data collected in the laboratory and then tested on a mutually exclusive subset of the same data. Results of these tests may significantly overestimate performance in the field. It could be that systems should be trained *and/or* tested on spontaneously-produced real user field data. The goal of this study was to evaluate the performance of a speaker independent isolated word telephone network speech recognition system when tested on laboratory vs. real user data under two training scenarios: 1. trained on laboratory and 2. trained on real user data. The results of this experiment suggest that real user speech databases are needed to achieve high accuracy speech recognition results in the field. In addition, it appears that a system trained on user data can be accurately tested with either real user or laboratory speech databases.

I. INTRODUCTION

1.1 Background

In developing and deploying automatic speech recognition systems, large amounts of speech data are required for system training and evaluation. While it is commonly accepted that the same acoustic data not be used for the two tasks, there is little beyond that in the literature to guide us in selecting training and testing databases.

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 speech produced by real users of the system under application-specific conditions. However, this is rarely done in practice. Instead, recognizers are most often trained and tested on laboratory speech data. A number of factors may be at play here. First, the collection of a real user field database is costly and time consuming. In the absence of data that quantifies the effects of training and/or testing on a laboratory vs. a real user database, it is not surprising that developers opt for the more cost effective solution. Moreover, if the application itself is under development, it may not be possible to collect a real user database. Finally, for the development of a general purpose recognizer, it may be undesirable to train (and perhaps to test) on an application-specific database.

As will be described in more detail below, laboratory and real user speech databases differ along two basic dimensions. First, they differ with respect to their 'source' characteristics; where source refers to the way the speech is produced. Second, they differ with respect to their 'transmission' characteristics; where transmission refers to the noise environment and the channel through which the data is recorded. A real user database is therefore defined as a recorded sample of speech that captures *both* the source and transmission characteristics of a target application.

If it were the case that speech recognition systems were performing as well in field applications as they were in the laboratory, the observations noted above would be of little interest. However, there is evidence that this is *not* the case. Recent data suggests that systems performing well in the laboratory often achieve significantly poorer results when faced with real user data [1].

The question under consideration in this paper is the following: Does the choice of laboratory vs. real user training and/or testing data have an effect on our ability to successfully deploy speech recognition technology in real applications? More specifically, how is performance affected by training on laboratory instead of real user speech? How much is the performance of a system overestimated by testing on laboratory rather than on real user data?

1.2 Real User vs. Laboratory Speech Databases

Real user databases are generally described as follows: casual users interacting with an automated or pseudo-automated system producing spontaneous goal-directed speech under application conditions. Depending on the specifics of the application, the database may reflect telephone transmission (e.g., from a home or public telephone) or wideband input (e.g., from an office, factory or public kiosk). These databases can be difficult and expensive to collect. In contrast, laboratory databases are described by some combination of the following characteristics: speech produced by paid speakers, acoustically-treated recording environments, high quality microphones, wideband communication channels, simulated or sampled telephone transmission channels, relatively few talkers and recited speech. Laboratory databases can be gathered relatively quickly and inexpensively. In addition, the characteristics of a laboratory database are relatively easy to control (in terms of speaker characteristics, noise backgrounds, etc.) .

The aforementioned features that differentiate real user from laboratory databases have not all been well-specified in terms of their acoustic consequences on the speech signals of interest. One differentiator that *has* received specific attention in the literature is that of spontaneously-produced vs. read speech.

A number of studies have investigated and documented perceptual and acoustic differences between spontaneously-produced and read speech. In these studies, each subject produces a spontaneous and a read version of the utterances under investigation. Remez et al. [2] found that when presented with sentence pairs, listeners can reliably distinguish between the spontaneous and read versions of the same utterance and that there appear to be multiple perceptual cues used to make the distinction. Zue et al. [3] showed that 1. spontaneously-produced sentences were longer in duration than their read counterparts (though

they found no difference in the overall distribution of phoneme durations), 2. spontaneously-produced sentences contained more pauses than their read counterparts and that the pauses were of longer duration and 3. spontaneously-produced sentences contained more 'non-speech vocalizations' (e.g., mouth clicks) and filled pauses than their read counterparts. Rudnicky et al. [4] showed that, at least for 1 out of 4 talkers, read utterances were produced at a *slower* speaking rate than their spontaneously-produced counterparts. These results are in contrast to those reported above by Zue. While it may be that this highlights the application-specific nature of these speech phenomena (i.e., Zue's application was an information and navigational assistance system while Rudnicky's application was a voice-driven spreadsheet system), more data is needed to establish the nature of this effect. Moreover, these two investigations did not purport to do an exhaustive study of the articulatory/acoustic differences between read and spontaneously-produced speech. It is well known that duration is not the only dimension along which the two types of speech differ (for example, see [2],[5]).

For each of these studies, the 'real user' speech was recorded under wideband application-like conditions. There have not been any direct acoustic or perceptual comparisons between laboratory and real user databases for telephone applications; 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. It may be that the differences reported above are similar for telephone speech. Alternatively, it may be that the acoustic/perceptual differences between real user and laboratory speech are larger and/or different for telephone than for wideband applications for the following reasons: 1. Source characteristics: When gathering real user data for the wideband studies described above, the talkers and the recognizer were in close proximity to one another and the talkers were aware that they were participating in a data collection experiment. When gathering real user speech for telephone applications, the talkers and the recognizer have been physically separated (linked only by a telecommunications channel) and the talkers have not been aware that they were participating in a data collection experiment (e.g., [1]), and 2. Transmission characteristics: During the collection of a wideband user speech database, the experimenter typically has control over the channel characteristics and noise backgrounds present during the recording sessions. This simplifies the distinction between user and laboratory speech databases; they differ mainly with respect to the source issue (i.e., the read-spontaneous speech issue). During the collection of a telephone user speech database, channel characteristics and noise backgrounds present during the recording sessions are out of the experimenter's control. They are likely to be extremely variable and not adequately modelled during the collection of the companion laboratory database; even if it is collected over telephone channels.

In summary, laboratory and real user speech databases differ along a number of important dimensions. There is wideband data to suggest acoustic and perceptual differences between the two database types; specifically along the read-spontaneous speech dimension. However, due to significant differences in data collection procedures, it is not clear that these results adequately represent either the source or transmission differences between laboratory and real user speech data for telephone applications.

1.3 Testing Databases

Our ability to specify the performance of speech recognition systems is important if we are to benchmark one system against another, determine the 'readiness' of a system for deployment, and estimate automation rates for a specific application. There are many open issues with respect to how performance evaluation should be done. The issue at hand here is to consider the effects of employing a real user vs. a laboratory-collected speech database for the purposes of performance assessment.

In the case of wideband applications, read and spontaneous utterances produced by the same talkers provide a straightforward way to address this issue. Jelinek et al. [6] compared the performance of a speech recognition system when tested on pre-recorded, read and spontaneous speech. Results indicate decreasing performance for the 3 sets of test material (98.0%, 96.9% and 94.3% correct, respectively). Rudnicky et al. [4] , on the other hand, evaluated their speech recognition system on both read and spontaneous speech 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 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. [7] also evaluated their speech recognition system on the read and spontaneous speech samples referred to above [3] and reported similar performance for the two data sets.

For the reasons addressed earlier, it has not been possible to collect databases that are matched with respect to speakers to consider the effects of laboratory vs. real user data 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].

1.4 Training Databases

It seems likely that a recognizer will perform optimally if trained on speech most like the target speech. However, there is little published data on the effects of laboratory vs. real user speech training databases in this context. Therefore, we cannot be sure of the size or even the presence of this effect.

With respect to wideband databases, Rudnicky et al. [4] claimed that since the performance of their recognition system (which was trained on read speech) was comparable on read and spontaneous testing data, the read vs. spontaneous distinction was irrelevant for training recognizers. More specifically, they claim that a system trained on read speech will not substantially degrade in accuracy when presented with spontaneous speech.

There is no published data on the effects of training speech recognition systems on laboratory vs. real user speech for telephone applications.

The goal of this study was to evaluate the effects of both the training and testing databases on the performance of a speech recognition system in order to address the following questions with respect to telephone speech recognition:

- What effect does training with real user vs. laboratory speech data have on overall system performance?
- How well can field performance predictions be made by testing speech recognition systems on laboratory data?

II. METHOD

2.1 Databases

Two isolated word databases were collected. Each database consisted of 179 instances of 15 New England city names produced over the telephone network (yielding two databases, each with 2685 utterances and approximately 15,900 phones).

2.2 Laboratory Database Collection

The laboratory speech database was produced by 179 talkers each of whom called a New York-based laboratory from their New England-based home or office telephone. Talkers were originally from the New England area and so were assumed to be familiar with the pronunciation of the target city names.

The laboratory data was collected using a Gradient Technology Inc. Desklab device connected via a SCSI bus to a Sun Microsystems SparcStation 1 workstation. The sampling frequency was set at 16000 samples/second, with 14 bits/sample. To reduce 'list effects', one of five randomized lists of city names was sent to potential participants. When a speaker called, the system answered the telephone and began playing digitized instructions. The talker was asked to speak his/her ID number (assigned by the agency recruiting the subjects), age, and sex, and then asked to speak the city names, waiting for a prompt before saying the next city name. The speech data was manually endpointed and orthographically transcribed.

2.3 Real User Database Collection

The real user speech database was produced by actual users of New England Telephone Directory Assistance. All users reached Directory Assistance by dialing '411' and through random assignment of calls to operator positions were assigned to the experimental position. Since each caller asked for a single city name, it seems reasonable to assume that the user speech database represents 2685 different speakers. It is also assumed that these New England callers knew the correct pronunciation of the city name they requested.

The user speech database was collected using an IEEE 488 relay and bus controller, an IBM PC/AT, a WORM optical disk drive for speech storage and a DIGI-SOUND device. User speech was sampled at 16000 samples/second, 16 bits/sample (see [8] for more detail). The recorded speech samples represent the users' responses to an automated request for a city name. Only isolated city name utterances were included in this study, representing a fraction of the total number of calls recorded. The speech data was manually endpointed and orthographically transcribed.

2.4 Experimental Design

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 [9]). The system was trained on each training set and then tested on each testing set[1]. This resulted in the following four training/testing conditions: 1. trained on laboratory speech and tested on laboratory speech, 2. trained on laboratory speech and tested on real user speech, 3. trained on real user speech and tested on real user speech and 4. trained on real user speech and tested on laboratory speech.

[1]For the laboratory database, the training and testing sets were mutually exclusive with respect to speaker.

In consideration of the widely held belief that more training data will always improve recognition performance, one final experimental combination was carried out. The system was trained on both laboratory *and* real user training data and then tested on each testing set. This added 2 further training/testing conditions to the above: 5. trained on both sets of data and tested on real user speech and 6. trained on both sets of data and tested on laboratory speech.

III. DATA ANALYSIS AND RESULTS

To compare our results to those of Zue and Rudnicky, the average duration of the read vs. spontaneously-produced city names was computed. Results indicate that the spontaneously-produced city names were slightly shorter in duration than their read counterparts (471 ms. and 525 ms., respectively). This agrees with the results from Rudnicky's paper, although the effect here is substantially smaller in terms of proportion of utterance length.

Percent correct scores were tabulated for each city name and submitted to a 2-way repeated measures analysis of variance with the factors: Training data (laboratory speech, user speech, both) and Testing Data (laboratory speech, user speech). Percent correct scores were submitted to a Logit transformation. While this data was used for the analysis, the figure shown here reflects actual percent correct scores. The analysis revealed a significant main effect of Training data, F (2,56) = 27.08, $p < .01$, a significant main effect of Testing data, $F(1,28) = 29.69$, $p < .01$, and a significant interaction between Training and Testing data, $F(2,56) = 52.85$, $p < .01$. Due to the significance of the interaction term, the main effects will not be discussed further.

Results are shown in Figure 1. As can be seen, performance changed little as a function of laboratory vs. user vs. both training databases when tested on laboratory speech (95.9 % vs. 91.1% vs. 94.8%, respectively). In contrast, performance changed dramatically as a function of training database when tested on user speech (52.0% vs. 87.0% vs. 87.4% for laboratory, user, and both training databases, respectively). To consider these results in more detail, a post hoc test of simple main effects was performed.

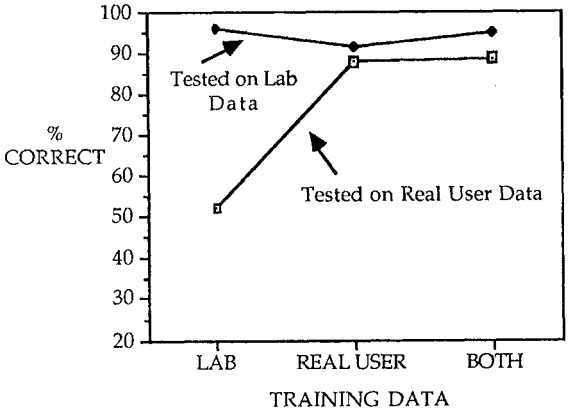

Fig. 1-The effects of Testing and Training databases

As expected, when testing on laboratory speech, the training database did not significantly affect recognition performance. In contrast, when testing on real user speech, the training database had a significant effect on recognition performance, $F(2,27) = 66.87$, $p < .01$.

Looking at a system trained on laboratory speech, performance was significantly affected by the test database, $F(1,28) = 221.98$ $p < .01$. That is, recognition results were significantly poorer when testing on user speech than when testing on laboratory speech. Looking at a system trained on real user speech, performance was not significantly affected by the test data. Finally, looking at a system trained on both laboratory and real user speech, there was no significant change in performance as a function of the test data.

IV. DISCUSSION AND CONCLUSIONS

The goal of this study was to evaluate the effects of both training and testing data on the performance of a speech recognition system for telephone applications.

The results of this study suggest that the selection of a training database can have a large effect on overall system performance for some conditions. More specifically, if a system is trained on laboratory speech, performance on laboratory speech will be high. However, when evaluating this system on real user speech, recognition accuracy will be substantially reduced. This highlights 2 issues:

- Field performance predictions cannot be accurately made by testing a laboratory-trained system on laboratory data. Such a test will yield misleading results. Many performance evaluations have been done this way in the past. It is easy to see why end users of such systems have been disappointed.
- It may not be feasible to train a system on laboratory data alone and realize high performance results on real user speech data; at least for this telecommunications application. To obtain high performance on real user data appears to require training data from that environment.

On the other hand, if a system has been trained on real user speech, system performance can be high, irrespective of the testing database. That is, an accurate assessment of field performance can be obtained by testing on user or on laboratory speech.

These results stand in contrast to those reported by Rudnicky et al. [4] and Zue et al. [7]. One possible explanation, as suggested in Section 1.2, might be that the differences between laboratory and real user data are substantially larger for telephone applications than for wideband applications. Assuming this to be true, it may not be necessary to collect 'real user' data for training a wideband speech recognition system (laboratory data should suffice) whereas it *does* appear to be necessary to collect real user data for training a telephone speech recognition system. An alternative explanation is that the data collection procedures used to collect spontaneous speech in the wideband studies cited above may not adequately model 'real user' data (as defined in Section 1.1) with respect to the source characteristics (i.e., the way typical users speak) and/or the transmission characteristics (i.e., the background noise present during typical use of the system). To clarify this issue, it would be useful to collect real user data (as defined here) for wideband applications.

Finally, the results of this study suggest that more training data is not necessarily better. When trained on both laboratory and real user data, recognition performance did not significantly improve for either test set. It may be that given a recognizer exposed to exemplars of the target speech, additional training data is only useful if it is the 'right data' (i.e., more examples of the target speech).

In considering this last result it is important to note that there are many uncontrolled differences between laboratory and user speech databases for telephone applications. In this experiment, for example, there were many more speakers for the user database than for the laboratory version. More data is currently being collected so that this disparity can be reduced and the results reconsidered. Nevertheless, most of the differences between the two databases used here accurately reflect inherent differences between typical laboratory and typical real user databases. Therefore, the results are relevant both to developers who are faced with a choice regarding training databases and to end users who depend on realistic predictions of field performance.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge the members of the MIT Spoken Language Systems Group who developed the base recognition system and assisted in various ways towards the completion of this experiment; Sara Basson, Maryann Cavan, Charles Jankowski, David Lubensky, John Mitchell, James Schrage, Erik Urdang, Dina Yashchin and the employees of New England Telephone Operator Services who made the collection of these databases possible; Sara Basson and Kim Silverman for their statistical consultations and editorial assistance.

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] Remez, R.E., Rubin, P.E., Nygaard, L.C. (1986) On Spontaneous Speech and Fluently Spoken Text: Production Differences and Perceptual Distinction, *Journal of the Acoustical Society of America*, Vol 79, Supp. 1, S26(A).

[3] Zue, V., Daly, N., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., Seneff, S. and Soclof, M. (1989) The Collection and Preliminary Analysis of a Spontaneous Speech Database, *Proceedings of the Second DARPA Speech and Natural Language Workshop*, October, 1989.

[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] Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., Miller, M. (1985) Measures of the Sentence Intonation of Read and Spontaneous Speech in American English, *Journal of the Acoustical Society of America*, 77, p. 649 - 657.

[6] 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.

[7] 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.

[8] Yashchin, D., Schrage, J. (1990) Data Acquisition Vehicle (DACQ) Design and Implementation, *NYNEX Science & Technology Technical Memorandum* (in progress).

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