ISSUES INVOLVED IN VOICEMAIL DATA COLLECTION

M. Padmanabhan, G. Ramaswamy, B. Ramabhadran, P. S. Gopalakrishnan,
C. Dunn *
IBM T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598

1 INTRODUCTION

Speech recognition is an important area of research today because of current trends in user interfaces and applications that advertise speech as an alternative to other forms of interacting with a computer. Initial attempts at developing practical speech recognition systems required the users to speak with pauses between words (isolated speech); subsequently with the advancement of speech recognition technology and increase in available computing power, the constraints on the users were relaxed to the point where the users could speak continuously but the performance of such systems was still less than satisfactory for conversational style disfluent speech. The goal today is to develop algorithms that will provide acceptable transcription accuracy with natural spontaneous disfluent speech.

Most speech recognition systems are based on observing the statistics of speech data in a training data set and generalizing it to a test data set. It is generally the case that to obtain good performance on a particular test set, it is necessary to train the system parameters on similar data - for instance one cannot expect to obtain good performance on spontaneous speech by training only on read speech data. Hence it is necessary to collect speech databases that contain the style of speech that we are trying to transcribe.

Naturally occurring speech in day-to-day life is almost exclusively continuous speech, and can further be broadly divided into two components, read speech and spontaneous speech. The latter category can further be broadly classified into the following classes:
(i) conversational monologues where one person is communicating with an audience with no feedback between audience and speaker (eg. radio broadcast news).
(ii) conversational monologues where one person is communicating with an audience but the person's speech is directed by feedback from the audience (eg. a seminar)
(iii) conversational interaction between a human and a machine with no feedback from machine to speaker (eg. voicemail)
(iv) conversational interaction between a human and a machine where feedback provided by the machine directs the speaker (eg. ATIS conversational systems)
(v) conversations between two speakers where feedback from each directs the others speech (eg. telephone conversations)
(vi) conversations between a number of people (eg. conference/teleconference).

Databases that currently exist for the purpose of conducting research in speech recognition include the Wall Steet Journal database which includes speech from approximately 300 speakers reading articles from the Wall Street Journal [1] (read speech). The Switchboard/CallHome database [2] (telephone conversation between two parties recorded at the switchboard, with the topic of the conversation either being prespecified (Switchboard) or unspecified (CallHome)) serves as a good example of category (v), and the Hub4 database [3] which represents recordings of radio broadcast news serves as a good example of category (i). Of the remaining categories, (iii) represents a large volume of data that one encounters in day to day life, and interactions of the form (iv) are also starting to become practical today. Consequently, there is a need to improve speech recognition performance on these categories of data by studying the characteristics of speech in these types of interactions. It is the goal of this paper to describe a methodology for collecting data in category (iii).

2 DATA COLLECTION METHODOLOGY

Voicemail data is unfortunately fairly difficult to collect because of privacy and legal issues. We adopted

[0]C. Dunn is with L&A Inc.

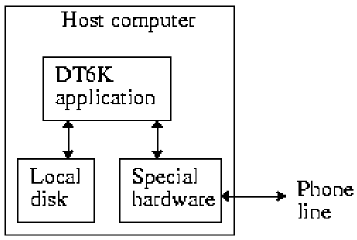

Figure 1:

the following strategy to build up a database. The data was collected at IBM sites at different locations in the US. Volunteers at these sites were asked to forward some of their voicemail messages to a local extension number (say 'abcd') set up for the purpose of collecting this data. The messages would then be collected periodically from the voicemailbox of this local extension and added to the database.

The volunteers were asked to forward only those messages that they felt were non-confidential and if neither they nor the person leaving the message had any objection to its being included in a database. Further, the volunteers were asked to add a sentence to their outgoing message of the form - 'Your voicemail may also be used for commercial research in developing algorithms for speech recognition. If you do not want your data to be used, please say so in your message.' - in order to let the caller know that their message may be included in the database, and to give them an opportunity to decline having their voicemail added to the database. Finally, the volunteers were provided with some incentives for every few messages that they forwarded.

The other aspect of the data collection procedure was to actually transfer the voicemail messages from the voicemail box of the local extension to the computer. In order to do this we used the DirectTalk6000 (DT6K) [4] software. DT6K is an application that runs under the AIX operating system on a host computer, and can interface to a phone line through special hardware on the host computer (see Fig. 1). The application provides a menu of functions to implement interactive voice response (IVR) applications; for instance, call a specified number, output a specified tone onto the phone line, record the audio data received on the phone line onto a disk on the host computer, etc. The application for collecting the data is embodied in the flow chart in Fig. 2. Note that the data was collected from IBM sites all over the US whereas the host computer that the DT6K application was running on was located at a single IBM site. Consequently, when

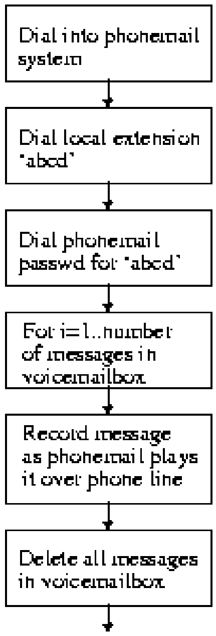

Figure 2:

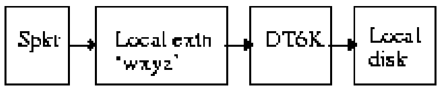

Figure 3:

the application dialed into the phonemail system of an IBM site in a different state, the voicemail messages were played out over a long distance line before they were recorded on the host computer.

The data was sampled at 8 KHz, and recorded in 8-bit μ-law compressed format onto a local disk of the host computer. Also as the messages were retrieved from the phonemail system, they had been compressed by the proprietary compression techniques used by the ROLM phonemail system, which is the phonemail system in use at various IBM locations.

2.1 Effect of phonemail compression on recognition performance

In order to evaluate the impact of the phonemail compression on speech recognition performance, we conducted some experiments. We collected data from three speakers (2 female, 1 male) reading some test sentences (sixty sentences related to business and office-correspondence) over the telephone. The setup for this experiment is shown in Fig. 3, i.e. the speaker

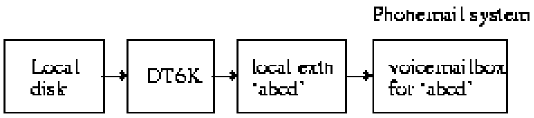

Figure 4:

would call the local extension number (wxyz) and any calls received at this number would be directly routed through to the DT6K application, which would then record the audio coming over the line onto the local disk. Note that this voice data is not subject to phonemail compression. The data was sampled at 8 KHz and stored in 8-bit μ-law format. This data was used to obtain a reference error rate for speech that has not been compressed by the phonemail system.

Subsequently, the data collected from these three speakers was played out again using DT6K to a different telephone extension 'abcd', where it was recorded by the phonemail system. This is shown in Fig. 4. The data from the three speakers is now stored as voicemail messages for the extension 'abcd'. This voicemail data (phonemail compressed data from the three speakers) was then retrieved from the phonemail system using the data collection setup of Fig. 2. The numbers in Table I give the word recognition error rates for the three test speakers based on using the uncompressed speech and using the phonemail compressed speech.

Table I

	Uncompressed	Compressed
Spkr 1 (m)	13.86 %	13.51 %
Spkr 2 (f)	16.94 %	16.23 %
Spkr 3 (f)	48.82 %	53.2 %

It can be seen from Table I that though the phonemail compression has some effect on the speech recognition error rate, it is generally small for most speakers.

3 TRANSCRIPTION OF TRAINING DATA

The voicemail training data collected from the phonemail system represents just speech data and it is necessary to get this transcribed before it can be used. The job of transcribing the data was subcontracted out to a transcription agency (which uses humans to listen to the speech and transcribe it). However, as this type of speech is typically extremely disfluent, the accuracy of the initial transcription was not very good. To calibrate the quality of the initial transcription, we computed the word error rate of the initial transcription of 78 messages (picked at random) and found it to be $\approx$ 9 % (we carefully handtranscribed these sentences and used these transcriptions as our reference in computing the error rate). This is fairly high and the use of the errorful transcriptions during the training process could lead to poorly estimated models. Consequently, we made a second pass through the training data and manually corrected the transcriptions.

The transcriptions obtained after this second pass still had inaccuracies. So we next devised an automatic scheme to identify possible transcription errors. This flagged around 1 % of the data, and we then corrected these transcriptions manually. The scheme was as follows: we first viterbi-aligned the voicemail data against the initial transcriptions using the baseline model. Subsequently, we computed the log-likelihood of each instance of a phone in the training data, conditioned on the alignment, and computed the average per-frame likelihood by normalizing by the number of frames that aligned to the phone. Then, we computed a histogram of these per-frame log-likelihood scores for each phone over all the training data. Next we went through the training data again and identified those instances of phones with per-frame likelihoods less than three 3σ below the mean per-frame likelihood for that phone (where σ represents the standard deviation of the score), and tagged the region of the acoustic corresponding to that instance of the phone as a possible transcription error. Finally, we listened to the tagged acoustic segments and manually corrected the transcriptions. Some examples of such corrections were

(i) we originally only had one baseform for IRA, AY AA R EY (the acronym baseform). In the recorded data IRA occurred as a name with pronunciation AY R AA, and was flagged as an error

(ii) there were several instances where disfluencies such as 'UH' and 'UM' had not been transcribed, and the technique flagged a number of these errors

The main objective in attempting this clean-up of the transcriptions was to obtain sharper acoustic models, and as the experimental results will show, this did help the error performance.

4 ANALYSIS OF VOICEMAIL DATA

At the time this paper was written, the Voicemail database comprised around 2200 voicemail messages, totaling 19.4 hours of speech. This data was collected from volunteers at various IBM sites in the US. The corresponding transcriptions of these messages had 220K words, and the size of the vocabulary corresponding

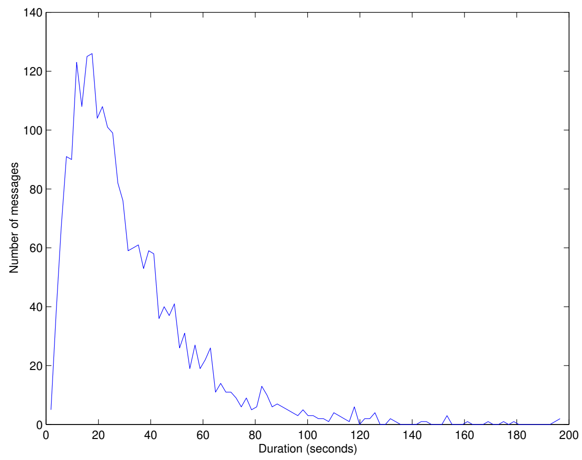

Figure 5: Histogram of length of messages

to these transcriptions was 9.7K words. We may conclude from these numbers that the average voicmail messages is 31 seconds in duration, and has about 100 words. These numbers are however misleading because we did encounter some very long messages (mainly related to technical subjects, for example a description of bugs in the latest software release of some application). Consequently, we plotted a histogram of the distribution of the lengths of these messages (see Fig. 5) and found that the peak of the histogram occurred at around 18 seconds.

The database was not entirely gender balanced because we did not do anything explicitly to ensure that this was the case. Approximately 38 % of the messages corresponded to male speakers.

We also did a subjective analysis of the topics covered by the messages, and found that the topics ranged from personal messages to extremely technically oriented messages. In this sense, the Voicemail database is again different from the Switchboard database where the speakers were asked to talk about a specified topic (one of 35 topics), and gives a distribution of topics in real-world voicemail. We attempted to subjectively characterize the topics into (a) business-related (eg. schedule for a meeting), (b)personal (eg. 'get back home before 9 pm ... or else' variety), (c) work-related (eg. maintenance schedule for a lab), (d) technical (messages contain technical information about programming), and (e) miscellaneous (messages not falling in any of the above categories). Based on a subjective categorization, we found that the percentage of these categories respectively was 27, 25, 17, 13 and 18 % respectively.

5 ACKNOWLEDGEMENT

We would like to acknowledge the support of DARPA for funding this data collection effort under Grant MDA972-97-C-0012. We are also extremely grateful to George di Simone and Ira Ellis (Watson telephon system support) for their help in setting up the data collection process. We would also like to thank Dr. Ellen Eide for helping with the verification of transcripts. Also, thanks are also due to Dr. Salim Roukos, Dr. David Nahamoo, and Dr. Lalit Bahl for their help and support. Finally, thanks are due to the various volunteers who contributed their voicemail messages to the database.

REFERENCES

[1] Proceedings of ARPA Speech and Natural Language Workshop, 1995, Morgan Kaufman Publishers.

[2] Proceedings of the LVCSR Workshop, Oct 1996.

[3] Proceedings of the DARPA Speech Recognition Workshop, Feb 1996, Morgam Kaufman Publishers.

[4] IBM AIX DirectTalk/6000 User Manual.