WSJCAM0: A BRITISH ENGLISH SPEECH CORPUS FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote and Steve Renals

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, U.K.

ABSTRACT

A significant new speech corpus of British English has been recorded at Cambridge University. Derived from the Wall Street Journal text corpus, WSJCAM0 constitutes one of the largest corpora of spoken British English currently in existence. It has been specifically designed for the construction and evaluation of speaker-independent speech recognition systems. The database consists of 140 speakers each speaking about 110 utterances. This paper describes the motivation for the corpus, the processes undertaken in its construction and the utilities needed as support tools. All utterance transcriptions have been verified and a phonetic dictionary has been developed to cover the training data and evaluation tasks. Two evaluation tasks have been defined using standard 5,000 word bigram and 20,000 word trigram language models. The paper concludes with comparative results on these tasks for British and American English.

1. INTRODUCTION

Appropriate databases are a fundamental necessity in the development of robust speech recognition systems. The increasing sophistication of recent systems demand sizable and well organised databases in their construction. Furthermore, such databases with associated tools provide a suitable environment for the performance evaluation of alternative recognition technology. This paper presents a new speech corpora for use in speech recognition that was designed with these needs in mind.

The speech corpus is intended to be the British English equivalent of the relevant parts of the North American English WSJ0 database [1]. The British English version consists of read speaker-independent (SI) material together with a verified transcription and automatically generated phoneme and word alignments. The corpus is partitioned into 92 training speakers, 20 development test speakers, and two sets of 14 evaluation test speakers. Each speaker provides approximately 90 utterances and an additional 18 adaptation utterances.

The resulting waveforms will be distributed by the Linguistic Data Consortium in compressed digitised form, accompanied by orthographic transcriptions, automatically generated phoneme and word alignments, a pronunciation dictionary and other relevant text material. The expected release date is early 1995.

Many other people have made a significant contribution to this work including Phil Woodland, Steve Young and Mike Hochberg.

This paper details the collection of this corpus and the associated information needed to use it. Comparative recognition results in British and American English are provided to demonstrate the utility of this resource.

2. SPEAKER CHARACTERISTICS

Speakers were recruited by placing advertisements throughout Cambridge University. These advertisements asked for native speakers of British English who would like to contribute to speech research by reading out newspaper sentences for an hour. Each speaker was offered a reward of £5 for their effort.

The selection of speakers was limited to people in and around Cambridge. However, effort was made to take advantage of the University's diverse population to find a wide range of regional accents. Figure 1 details the sex/age range distribution of the training set speakers, the figures for the development and evaluation test sets reflect the same distribution.

Range	18-23	24-28	29-40	> 40
Female	21	11	3	4
Male	25	19	4	5

Figure 1: Age Range Distribution of Training Speakers

3. RECORDING CONDITIONS

Recordings were made in an acoustically isolated room in Cambridge University Engineering Department. This "quiet room" measures five by five meters, and is acoustically isolated by double entry doors and double-glazed windows. In the room was a desk with a workstation monitor, keyboard, and mouse as well as a far-field desk microphone and a preamplifier for the head-mounted microphone. To circumvent the problem of fan noise, the actual workstation was located outside the room.

Recordings were made with both a Sennheiser HMD 414 head-mounted microphone and an inexpensive Canford desk microphone. The desk microphone was positioned about 1/2 meter from the speaker's head, approximately thirty degrees to the speaker's left at desk level. A Symmetrix SX-202 preamplifier was used to convert the close-talking microphone signal to line level; the desk microphone

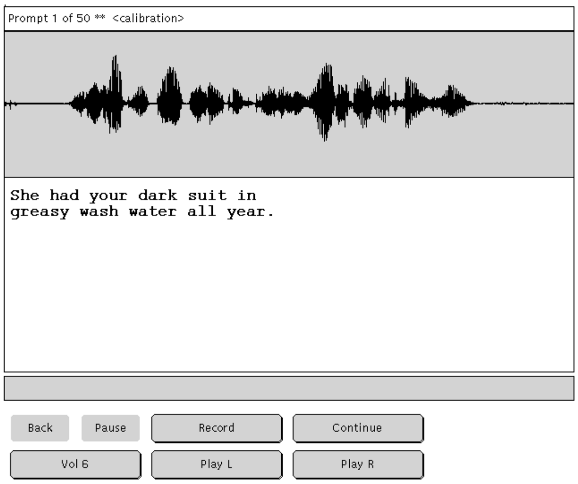

Figure 2: Data capture user interface

was amplified by a custom circuitry. Both microphone signals were digitised at 16kHz by the A/D converters internal to a Silicon Graphics Iris Indigo workstation. The two (independent) microphone signals were recorded simultaneously as a stereo signal for exact time-alignment.

The background noise level was measured as 28 dB(A). The average SNR was 35–45dB for the close-talking microphone and 20–25dB. for the desk microphone (computed using the NIST SPHERE utilities 'speech' and 'segsnr'). Despite the lower SNR, recordings made with the cardioid desk microphone are subjectively crisp, with little of the audible reverberation characteristic of omnidirectional microphones.

4. RECORDING PROMPTS

Recording sessions started with the speaker reading instructions on paper. These instructions request a natural style of speaking and warn of the technical nature of the prompt material. The technical instructions on how to operate the recording software were clarified by a short demonstration, after which speakers had the opportunity to ask further questions.

The actual recording process for each sentence consisted of five steps. First, speakers had to read through the sentence in silence. Then they clicked a **record** button on screen and read the sentence out aloud. After they had finished recording they clicked the **stop** button. They could subsequently click **play** to play back their recording. If they were pleased with the current result they clicked **continue** to go to the next sentence. If not, they clicked **record** again to re-record the sentence, which they could do as many times as they liked. They were required to only play back and correct sentences if they weren't absolutely confident about the correctness of their utterance.

After a supervised practice session on four sentences, any problems the speaker had were explained and the recording level was adjusted accordingly. Subsequently, the real recording started with the 18 adaptation utterances. The co-ordinator then left the room before the recording of either 90 training sentences or 80 test sentences. Finally, a short anonymous questionnaire was completed.

All recorded sentences were taken from the Wall Street Journal (**WSJ**) text corpus. This large corpus of sentences was selected from the 1987-89 editions of the US business newspaper. Since this corpus had previously been recorded in US American English for identical purposes, the existing recording experience was at our disposal [1]. The advantage of using WSJ thus was that we could make use of widely available existing conventions, utilities, vocabularies, and large selections of processed texts from a real newspaper. The main problems with using WSJ sentences in eliciting British English utterances arose from their U.S. origin. This posed an extra pronunciation problem to some speakers, which came on top of problems that some speakers had with WSJ's financial jargon and typical style of phrasing.

5. CORPUS STRUCTURE

The recording text material consisted of adaptation, training and testing components. The same set of 18 adaptation sentences was recorded by each speaker, consisting of one recording of background noise, 2 phonetically balanced sentences and the first 15 adaptation sentences from the initial WSJ experiment.

The *training* sentences were taken from the WSJ0 training subcorpus of about 10,000 sentences. Each training speaker read out some 90 training sentences, selected randomly in paragraph units. This is the empirically determined maximum number of sentences that could be squeezed into one hour of speaker time. The same sentences were allowed to occur in several speakers' prompts, though never more than once in the same speaker's.

Each of 48 *test* speakers read 80 sentences from the subcorpus originally designated for development testing in WSJ0, consisting of 40 sentences from the 5,000 word corpus (2,000 sentences) and 40 sentences from the 64,000 word corpus (4,000 sentences). These sentences were randomly selected and spoken by no more than one speaker. The test material was recorded without deciding a priori on a division of speakers into development and 2 evaluation speaker groups. A balanced distribution of test speakers over development and evaluation sets was obtained on the basis of gender, age group and lastly approximate speaking rate. The final development test group consists of 20 speakers which are to be distributed with the training material. Each of the two evaluation test groups comprise of 14 speakers and will be released at a later date.

6. PRONUNCIATION DICTIONARY

A British English Example Pronunciation Dictionary (BEEP) has recently been developed for large vocabulary speech recognition with the WSJCAM0 corpus in mind. The main body of this pronunciation dictionary was derived from the Computer Usable Version of the Oxford Advanced Learner's Dictionary (CUVOALD)[1] and the MRC Psycholinguistic Database[2]. These covered 65% of the 22,000 pronunciations needed for this task. Approximately 1,000 of the remainder were automatically derived, 5,000 were entered by

[1] Available from
`ftp://black.ox.ac.uk/pub/ota/dicts/710/`

[2] Available from
`ftp://black.ox.ac.uk/pub/ota/dicts/1054/`

hand and 2,000 were taken from the CMU dictionary[3] of American pronunciations.

The phone set used was an extension of the ARPA-bet symbols used in the CMU dictionary and the TIMIT database. Three new symbols were needed to cover British English, these were named /oh/ for the vowel in "pot", and /ia/, /ea/ and /ua/ for the diphthongs in "peer", "pair" and "poor" respectively. The resulting dictionary covers 90,000 words and is available for non-commercial use by Internet FTP from host `svr-ftp.eng.cam.ac.uk` with file name `/pub/comp.speech/data/beep-0.2.tar.Z`.

7. EVALUATION TASKS

The development and evaluation test data is partitioned into two tasks, a 5,000 word closed vocabulary task, and an open vocabulary task which in practice has a 64,000 word vocabulary. The development test data is further divided by filtering out utterances where there was a problem in the recording and splitting the remainder into two. This yields four sets of development test utterances, named `si_dt5a`, `si_dt5b`, `si_dt20a` and `si_dt20b` with about 360 utterances in each set.

Standard ARPA CSR evaluation language models may be used with these vocabularies, such as those provided by MIT Lincoln Laboratories[4]. The standard language models are the 5,000 word bigram and 20,000 word trigram models, and results for these tasks on `si_dt5a` and `si_dt20a` are reported below.

8. SPEECH RECOGNITION RESULTS

We have built a complete British English continuous speech recognition system using the WSJCAM0 database. We used the complete WSJCAM0 training set of around 8,000 utterances, excluding the adaptation utterances. We tested on the `si_dt5a` and `si_dt20a` test sets using the standard 5k closed and 20k open language models respectively. We used the ABBOT large vocabulary continuous speech recognition system [2] for this work. ABBOT is a high-performance hybrid recurrent network/HMM system, that has been used in the ARPA SLT evaluations. We used the BEEP dictionary and the Nov 1993 standard 5k bigram and 20k trigram backed-off language models. Our recognition results for WSJCAM0 are shown in table 1. These results compare well with the equivalent system trained on the ARPA WSJ database of similar size and character (WSJ0, SI-84), when tested on the November 1993 evaluation data (table 2).

These WSJCAM0 results were obtained using a single recurrent network acoustic model, with a front-end based on a Mel-scaled filter bank. We expect these results to be substantially improved when we use the full ABBOT acoustic model, which involves the combination of four recurrent networks trained using different front ends. Table 2 illustrates the improvement on the ARPA WSJ task.

9. CONCLUSION

This paper has presented a new speech corpus of spoken British English which is sufficient for training large vocabulary speech recognition systems. Existing formats and protocols as defined by the ARPA CSR community have been followed as closely as possible. This has been done to minimise the effort needed to build a British English recognition system. Such a system has been built and the result is one of the largest continuous speech recognition systems that have been constructed to date.

[3] Available from `ftp://ftp.cs.cmu.edu/project/fgdata/dict/`

[4] Available from `ftp://where?`

WSJCAM0					
Task	Correct	Subst.	Del.	Ins.	Error
5K bigram	86.5	10.4	3.1	1.6	15.1
20K trigram	80.6	16.2	3.2	3.5	22.8

Table 1: Results on WSJCAM0 5k closed and 20k open development test sets, using standard backed-off bigram and trigram models respectively. The acoustic model used a single recurrent network with a Mel+ front end.

WSJ0, SI-84 training, Nov 93 Eval Data		
Task	Single Network Error	Combined Networks Error
5K bigram	16.9	11.4
20K trigram	22.4	18.2

Table 2: Results on WSJ0, 5k closed and 20k open Nov 93 evaluation sets, using standard backed-off bigram and trigram models respectively. The single network system used a Mel+ front end, the combined system averaged the outputs of 4 recurrent networks trained on 4 different front end processings of the same data.

10. ACKNOWLEDGEMENTS

This work was supported in part by LRE project 62-058, SQALE, DTI/EPSRC Grant Ref IED4/1/5804 and the Linguistic Data Consortium. Two of the authors, T.R. and S.R., hold U.K. Engineering and Physical Sciences Research Council Fellowships. The pronunciation dictionary for the American English system was provided by Dragon Systems, and the language models were provided by MIT Lincoln Laboratories.

11. REFERENCES

[1] Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc. Fifth DARPA Speech and Natural Language Workshop*, pages 357–362. DARPA, Morgan Kaufman Publishers, Inc., 1992.

[2] Mike Hochberg, Tony Robinson, and Steve Renals. Large vocabulary continuous speech recognition using a hybrid connectionist HMM system. In *Proceedings of the International Conference on Spoken Language Processing*, 1994.