SVitchboard 1: Small Vocabulary Tasks from Switchboard 1

Simon King[*]

Centre for Speech Technology Research
University of Edinburgh, UK
www.cstr.ed.ac.uk

Chris Bartels, Jeff Bilmes[†]

Department of Electrical Engineering
University of Washington, USA
ssli.ee.washington.edu

Abstract

We present a conversational telephone speech data set designed to support research on novel acoustic models. Small vocabulary tasks from 10 words up to 500 words are defined using subsets of the Switchboard-1 corpus; each task has a completely closed vocabulary (an OOV rate of 0%). We justify the need for these tasks, describe the algorithm for selecting them from a large corpus, give a statistical analysis of the data and present baseline whole-word hidden Markov model recognition results. The goal of the paper is to define a common data set and to encourage other researchers to use it.

1. Introduction

Currently, one of the most challenging tasks in speech processing is the recognition of spontaneous conversational speech. Compared to carefully read speech, conversational speech has more variation in pronunciation and speaking rate as well as a higher degree of coarticulation. Another difficulty in studying spontaneous conversational speech is that, although the core vocabulary of everyday speech may be just a few thousand words, the possible vocabulary is typically in the tens or hundreds of thousands of words: so there are a large number of tokens of low-frequency words.

Researchers attempting to build better models of the coarticulation and pronunciation variation present in spontaneous speech are currently faced with the choice between two unsatisfactory extremes: small vocabulary read speech corpora such as OGI Numbers[1] and (very) large vocabulary spontaneous speech corpora such as Switchboard [1] and the like.

1.1. Motivation

Working with computationally expensive models (e.g. articulatory approaches such as [2, 3, 4, 5, 6]) presents problems. Near-real-time and limited memory first-pass decoding on large vocabulary corpora using complex models is currently impossible. On the other hand, working with small vocabulary corpora is quite feasible, but these corpora are usually not of spontaneous or conversational speech.

It is useful to be able to decouple the problems of developing a novel acoustic model from the problems of: constructing a lexicon; language modelling; decoding; dealing with words that are unseen in the training data. A carefully designed closed-vocabulary task could eliminate or avoid these problems.

The standard solution is lattice rescoring; this is not entirely satisfactory. A large lattice does not sufficiently limit computation. The low word error rate (WER) region of search space represented by a small lattice may not overlap with the low WER region of a novel model, thus preventing the novel model achieving a low WER by rescoring this part of search space. First-pass decoding is preferable and makes error analysis easier (i.e. recognition errors are entirely due to the novel model, not some combination of novel model and lattice).

An alternative solution, which we present here, is a corpus of spontaneous, conversational speech with a small, limited vocabulary. Of course, such a corpus does not exist and could not be recorded. We introduce *SVitchboard 1*, the Small Vocabulary Switchboard 1 database, which is composed of selected utterances from the Switchboard I corpus [1]. SVitchboard 1 contains a number of tasks with increasing vocabulary sizes: 10, 25, 50, 100, 250 and 500 words.

This paper will begin by describing the algorithm used to construct the database in section 2. Section 3 defines a five-fold cross-validation procedure. We give statistical information about the resulting corpus in section 4 and section 5 gives baseline recognition results, which we invite other researchers to beat.

2. Creating the corpus

All of the data in SVitchboard 1 are taken from the Switchboard 1 corpus of two-person telephone conversations [1]. We present a simple algorithm for simultaneously finding a vocabulary of any given size and selecting the corresponding in-vocabulary utterances from the larger cor-

[*] Supported by EPSRC Advanced Research Fellowship GR/T04649/01. Work partially carried out whilst at the University of Washington, funded by EPSRC grant GR/T12118/01.

[†] Partially supported by NSF grant IIS-0093430 and an Intel Corporation Grant.

[1] http://www.cslu.ogi.edu/corpora/numbers/index.html

Task	Partition	Utterances	Word tokens	Duration (hours)	
				Total	Speech
10	A	1384	1617	0.67	0.20
	B	1275	1455	0.60	0.17
	C	1196	1389	0.56	0.16
	D	1446	1628	0.69	0.20
	E	1474	1703	0.70	0.20
	Total	6775	7792	3.22	0.93
25	A	1943	2698	0.95	0.29
	B	1887	2560	0.90	0.26
	C	1732	2359	0.83	0.25
	D	2078	2789	1.01	0.30
	E	2138	2918	1.04	0.31
	Total	9778	13324	4.74	1.42
50	A	2474	4228	1.24	0.39
	B	2392	3932	1.16	0.36
	C	2233	3789	1.10	0.34
	D	2594	4292	1.29	0.40
	E	2749	4673	1.37	0.43
	Total	12442	20914	6.16	1.93
100	A	2916	5814	1.51	0.51
	B	2794	5290	1.40	0.46
	C	2632	5237	1.34	0.45
	D	3059	5981	1.57	0.53
	E	3201	6289	1.64	0.55
	Total	14602	28611	7.47	2.48
250	A	3741	10400	2.10	0.81
	B	3681	10060	2.02	0.77
	C	3415	9336	1.88	0.71
	D	3927	10581	2.18	0.83
	E	4169	11573	2.32	0.89
	Total	18933	51950	10.50	4.01
500	A	4675	17948	2.92	1.30
	B	4673	17519	2.86	1.26
	C	4249	15857	2.60	1.13
	D	4871	18075	3.00	1.32
	E	5202	20021	3.23	1.43
	Total	23670	89420	14.62	6.44

Table 1: Data set sizes

pus. The desirable properties of the limited-vocabulary corpus being created include:

Closed vocabulary: there is a fixed, known vocabulary and that every utterance in both train and test sets only contains words from within this vocabulary.

Balance: all words should have examples in training, validation and test sets. If words (or other units) in the test set do not occur in the training set, then whole-word (or other unit) modelling is simply not possible. Whole word modelling is useful in articulatory models, for example, where one wishes to avoid phone-sized units altogether.

Minimal number of low-frequency words: the long tail of singletons and low-frequency word types in the typical Zipf-like distribution of a corpus like Switchboard 1 is a problem for novel approaches.

Maximum size: the maximum amount of data (number of word tokens) should be selected.

The first step in creating the limited vocabulary corpus was to divide each side of the long conversations of Switchboard 1 into shorter segments. The initial cuttings used here were those published by Mississippi State University [7]. To maximise the amount of in-vocabulary utterances available for selection, we further cut these segments into smaller utterances at every silence longer than 500ms. This algorithm is then used:

sv_vocabulary = 5 most common words in large corpus
oov_vocabulary = full_vocabulary $\setminus$ sv_vocabulary
while |sv_vocabulary| $<$ target number of words **do**
  for all $word \in$ oov_vocabulary **do**
    new_vocabulary = sv_vocabulary $\cup\ word$
    incoming_utterances = all utterances that only contain words in new_vocabulary
    count$_{word}$ = number of words in incoming_utterances
  end for
  $new_word = \arg\max_{word}$ count$_{word}$
  sv_vocabulary = sv_vocabulary $\cup\ new_word$
  oov_vocabulary = oov_vocabulary $\setminus\ new_word$
end while

Labelled silences are always allowed, but do not contribute to count$_{word}$. The algorithm incrementally adds one word at a time to the vocabulary. Larger vocabularies can be built by initialising sv_vocabulary to a previously found smaller vocabulary. We decided on vocabulary sizes of 10, 25, 50 ,100, 250 and 500 words (*plus* the pseudo-word `sil`), which gives users a very simple task to start with (e.g. for debugging their model) and a sequence of increasingly difficult tasks to work through. The vocabulary of each task is a subset of all larger ones: e.g. all words in the 10 word task also appear in all the larger tasks. The speech data for each task is likewise a subset of the data in all larger tasks. The 10 word vocabulary (in decreasing order of frequency) is: right, oh, okay, so, well, and, yes, really, I, the.

2.1. Dealing with disfluencies and other problem words

Some transcribed words or pseudo-words in Switchboard 1 present problems for various reasons, including extreme variation in pronunciation or difficulty in writing pronunciations for them. Since SVitchboard 1 is designed to be useful for novel *acoustic* modelling work, we decided to simplify the corpus by excluding utterances containing any of these words: all word fragments (e.g. sim[ilar]-), words ending in a digit, uh, [noise], i-, yeah, [laughter], huh, hm, [laughter-*], uh-huh, um-hum hum, huh-uh, um. The tasks defined in this paper should be referred to as "SVitchboard 1: no filled pause condition". A version in which filled pauses are allowed may be constructed in future.

3. Cross-validation procedure

The speakers (and their speech data) are divided into non-overlapping subsets to create a 5-fold, speaker-independent, cross-validation scheme. Five *partitions* were constructed, denoted by the letters A to E, each of 108 speakers (partitions A–C) or 107 speakers (partitions D and E). For any given limited vocabulary task (e.g. the 50 word task), we take only the in-vocabulary utterances from each parti-

Subtask	Train	Validate	Test
1	ABC	D	E
2	BCD	E	A
3	CDE	A	B
4	DEA	B	C
5	EAB	C	D

Table 2: Definition of the five-fold cross-validation setup. Numbers 1–5 denote the five subtasks, and letters A–E denote the five partitions of the data.

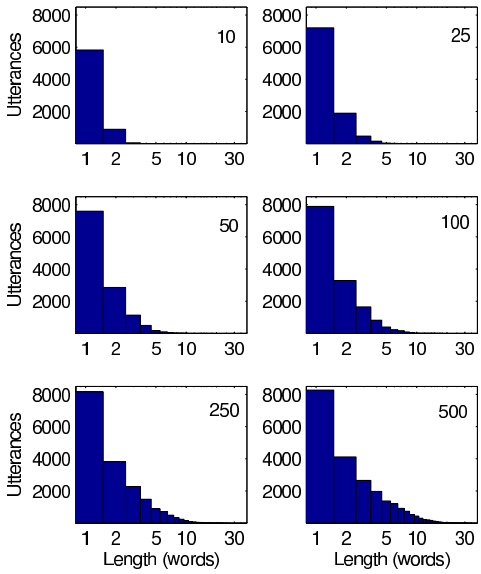

Figure 1: Utterance length distributions (horizontal axis is logarithmically scaled).

tion.

Now we can use these partitions to define five *subtasks* using a jackknife procedure, as shown in table 2. In each subtask, three partitions from {A, B, C, D, E} are formed into a training set, with the remaining two sets being used as validation and test sets respectively.

3.1. How to use the corpus

Each vocabulary size corresponds to a *task*, which should be referred to as, for example, "SVitchboard 1: no filled pause condition, 50 word task". Within each *task* there are five *subtasks*, numbered 1–5. Each *subtask* specifies a particular arrangement of the five *partitions* (labelled A–E) into training, validation and testing sets.

Across the 5 subtasks, each partition appears the same number of times in the training set, in the validation set and in the test set (table 2). Thus any inter-partition variation in recognition difficulty is cancelled out if overall results are reported across all 5 tasks (as in table 3).

Each subtask is to be performed independently of the others. For computationally expensive systems, the first subtask can be used alone. If training is particularly expensive, then only the first partition of the training set can be used (e.g. in subtask 1, train on A only, rather than A, B and C), but this should be done only as a last resort. Since one of the aims of this corpus is to facilitate direct

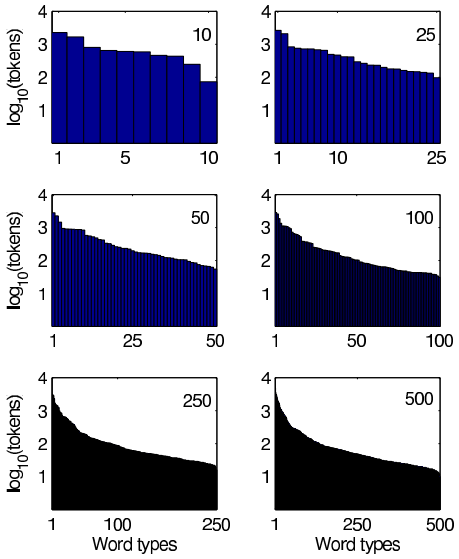

Figure 2: Word frequency distributions have a truncated Zipf shape.

comparison of systems on identical data, the training set should not be enhanced (e.g. by adding Fisher data).

Where practical, it is preferable to perform a set of five independent experiments (e.g. as is done in section 5 of this paper) and report results both by subtask and as an overall word error rate per task. For systems employing a language model, this model must also be trained on the training data specified in table 2. That is, the language model cannot be shared across all five experiments, unless it is trained on data other than Switchboard 1 transcripts.

Phonetically labelled data: partition E includes the 12 speakers for which there is phonetically labelled data from the ICSI Switchboard Transcription Project [8], and the remaining speakers were divided randomly. The phonetically transcribed data could be used either as a reliable reference for phone recognition (on either validation or test data) or as data to be used in co-training.

Split conversations: if a language model is trained on one side of a conversation and used to decode data taken from the other side of the conversation, this has the potential to give the language model an unfair advantage. This effect is reduced in SVitchboard 1, since low-frequency topic words (likely to be spoken by both speakers in a single conversations) are less likely to be in the limited vocabularies. However, it should be noted that, in SVitchboard 1, it is common to find one side of a conversation in the training set and the other other side in the corresponding test set. It is impossible to avoid this, since any given speaker will have conversations with multiple other speakers (who will have conversations with further speakers...); it is impossible to ensure both sides of all those conversations appear within a single partition.

Cross-talk: Note that a significant amount of cross-talk can be heard, especially during silence portions.

4. Statistical analysis of the corpus

The sizes of the data sets can be seen in table 1. Utterance length distributions are shown in figure 1 and the word frequency distributions are in figure 2. In all but the 500 word task, every vocabulary word appears in every partition (A–E) and therefore in every train, validation and test set. In the 500 word task, each partition has between 1 and 4 words fewer than 500, but the missing words are different for every partition so therefore the subtask training sets (ABC, BCD, etc) all contain every vocabulary word. The lowest frequency word in each subtask (adding all 5 partitions together) has a frequency of 73, 97, 55, 31, 16, 10 in the 10, 25, 50, 100, 250 and 500 word tasks, respectively.

5. Baseline results

A simple baseline whole word HMM system was built using HTK [9]; the number of states was 3× the number of phones for that word entry in the lexicon distributed with [7]. Observation vectors were 12 MFCCs plus energy, and their deltas and accelerations. Gaussian mixture distributions were trained using the HTK "mixing-up" procedure and the number of components was chosen per subtask to maximise validation set accuracy. Simple bigram language models were constructed per subtask using the HTK program `HLStats` with a bigram count threshold of 3; perplexities are shown in table 3. The word insertion penalty and language model scaling factor were optimised per subtask to maximise accuracy on the validation set. Results are shown in table 3.

6. Conclusion

SVitchboard 1 allows researchers to decouple the problems of developing novel acoustic models from the various problems of dealing with a large vocabulary. Future work will include publishing tasks with larger, but still limited, vocabulary sizes and also tasks with disfluencies (filled pauses, word fragments). Thanks are due to Mississippi State University for their freely available word alignments. SVitchboard 1 can be downloaded from http://www.cstr.ed.ac.uk/research/projects/svitchboard.

Task	Subtask	Perplexity		WER(%)	
		Val	Test	Val	Test
10	1	3.1	3.2	20.2	20.8
	2	3.2	3.4	20.1	21.3
	3	3.4	3.4	21.6	21.0
	4	3.4	3.3	21.3	24.5
	5	3.3	3.1	23.3	20.9
	overall				21.6
25	1	5.0	5.2	35.6	34.9
	2	5.2	5.3	35.5	35.9
	3	5.2	5.4	35.0	37.2
	4	5.4	5.3	35.4	37.9
	5	5.2	4.9	37.6	35.4
	overall				36.2
50	1	7.6	8.1	48.4	48.4
	2	8.1	8.1	48.6	46.3
	3	8.1	8.4	46.1	49.3
	4	8.3	8.1	48.1	51.4
	5	8.2	7.6	50.8	48.3
	overall				48.7
100	1	11.4	11.5	57.9	56.8
	2	11.4	11.7	56.4	55.1
	3	11.7	11.7	55.1	58.8
	4	11.6	11.6	55.6	57.4
	5	11.8	11.5	57.5	57.2
	overall				57.0
250	1	21.7	23.2	65.8	66.2
	2	23.2	22.3	66.8	64.4
	3	22.2	23.5	63.9	67.1
	4	23.4	22.4	65.3	66.7
	5	22.6	21.7	66.7	64.5
	overall				65.7
500	1	38.4	39.5	69.8	70.8
	2	39.4	38.5	70.0	67.9
	3	38.1	39.7	67.6	70.0
	4	39.4	38.0	69.2	69.7
	5	38.2	37.9	70.1	68.9
	overall				69.5

Table 3: Baseline LM perplexities, word error rates.

7. References

[1] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, California, March 1992, vol. 1, pp. 517–520.

[2] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 33, no. 2-3, pp. 93–111, Aug 1997.

[3] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-Articulator Markov Models for Speech Recognition," *Speech Communications*, vol. 41, no. 2, October 2003.

[4] Y. Zhang, Q. Diao, S. Huang, W. Hu, C. Bartels, and J. Bilmes, "DBN based multi-stream models for speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, China, April 2003.

[5] M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic Bayesian networks," in *Proc. IEICI Beyond HMM Workshop*, Kyoto, Dec. 2004.

[6] K. Livescu and J. Glass, "Feature-based pronunciation modeling for speech recognition," in *Proc. HLT/NAACL*, Boston, Massachusetts, May 2004.

[7] N. Ganapathiraju Deshmukh, A. Gleeson, A. Hamakera, and J. Picone, "Resegmentation of SWITCHBOARD," in *Proceedings of the International Conference Spoken Language Processing*, 1998, vol. 4, p. 1543.

[8] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *In Proc ICSLP*, Philadelphia, 1996.

[9] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland, *HTK manual*, Cambridge University Engineering Department, 2002.