RESEGMENTATION AND TRANSCRIPTION OF SWITCHBOARD

Jonathan Hamaker, Neeraj Deshmukh, Aravind Ganapathiraju, Joseph Picone

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University, Mississippi State, Mississippi 39762
{hamaker, deshmukh, ganapath, picone}@isip.msstate.edu

ABSTRACT

The SWITCHBOARD (SWB) Corpus has become one of the most important benchmarks for assessing improvements in large vocabulary conversational speech (LVCSR). The high error rates on SWB are largely attributable to an acoustic model mismatch, the high frequency of poorly articulated monosyllabic words, and large variations in pronunciations. It has been seen that an improved quality of segmentations and transcriptions translates well to improved acoustic modeling. The goal of our SWB resegmentation project is to resegment the data into utterances of approximately 10 secs. in duration using boundaries based on naturally-occurring silence, and to correct the transcriptions. A system trained on a subset of this data resulted in a 1.9% absolute reduction in word error rate. Equally exciting is the fact that recognition error rates on monosyllabic words dropped from 70.0% to 63.3% — a decrease of 6.7%. Since monosyllabic words dominate the SWB corpus, this is a particularly significant result.

1. INTRODUCTION

One of the most challenging tasks for state-of-the-art LVCSR systems is to accurately recognize conversations over the telephone. The SWITCHBOARD (SWB) Corpus [1] is currently used as a standard benchmark for such applications. The database contains 2430 conversations averaging 6 minutes in length. SWB totals over 240 hours of recorded speech, and about 3 million words of text, spoken by over 500 speakers of both sexes from every major dialect of American English.

The difficulties in recognition arise from short words, telephone channel degradation, and disfluent and coarticulated speech typical of casual conversations. Fast speaking rates, a wide range of dialects, speaking styles and accents, and the large variation in pronunciations of words all present unique problems for recognition of such spontaneous speech. These conversations are heavily populated with dysfluencies such as ungrammatical pauses, stutters, laughter, repeats and self-repairs. The vocabulary is large and dominated by monosyllabic words which are typically hard to recognize.

Analysis of current recognition performance on SWB attributes a significant proportion of the word error rate (WER) to monosyllabic words [2]. This is partly because monosyllabic words dominate the database, but more significantly because the transcriptions for such words are frequently in error. Hence, we believe that an important step in improving overall performance on SWB is to improve the quality of the training database. We have earlier demonstrated that an improved transcription of the test database results in significant (over 2%) improvement in the absolute WER [2]. Thus, there is definitely merit to resegmenting SWB and retraining acoustic models with cleaner transcriptions.

2. SEGMENTATION OF SWITCHBOARD

Segmentation of conversational speech into relatively short phrases enhances the transcription accuracy, helps in reducing the computational requirements for training and testing each utterance, and simplifies the application of the language model (LM) during recognition.

Linguistic segmentation is effective in maintaining clear linguistic context, but it has two significant problems. First, if the boundaries are based solely on language rules, boundaries may be placed between words where there is little or no silence. This will result in word beginnings and ends being cut off, which adversely effects training of acoustic models. Second, linguistically based boundaries often result in utterances which are too long for practical recognition systems. Speakers in SWB sometimes

carry on monologues of the same thought for 30-60 seconds, but the ideal utterance length for LVCSR systems needs to be closer to 10 seconds to limit excessive use of computational resources.

Segmenting speech based solely on acoustic boundaries also has its pros and cons. It is a more desirable paradigm in that boundaries are only placed where there is a pause, but this method obscures any inherent linguistic context. Thus, it is of little use when training language models. We try to achieve a tradeoff between the two paradigms by — manually placing boundaries where there is acoustical silence, maintaining linguistic context, and regulating the length of the utterances.

The resegmentation procedure is as follows:

- echo cancellation [5];
- manual adjustment of utterance boundaries;
- correction of the orthographic transcription of the new utterance;
- readjustment of boundaries if necessary;
- supervised recognition on the new utterances to get a time-aligned transcription;
- review of the word boundaries and final correction of transcriptions.

The challenging part of the correction process is the decision on whether to split at natural linguistic boundaries (sentence boundaries, turn boundaries, phrase boundaries, etc.) or at acoustical boundaries where there is a pause. Our strategy for resegmentation is as follows:

- segment at locations where there is clear silence separating each segment (at least 1 second long);
- segment along phrase, sentence, and/or train-of-thought boundaries.

The first rule is important because it eliminates the problem of truncated words due to segment boundaries falling where there was not enough separation between words. The second rule is formulated to maintain linguistic context and clarity for speech understanding and language modeling experimentation. We have modified these general guidelines to be specific and easily implemented by our transcribers:

- set boundaries so that each utterance has a beginning and ending silence of at least 0.5 seconds;
- utterances should be split to be approximately 10 seconds in length.

Thus the new manual segmentation of the training database consists of utterances typically less than 10 seconds in duration which are excised at significant pause boundaries and/or turn boundaries. Besides adjusting the utterance segment boundaries, we also correct various transcription errors such as typographical mistakes, inserted skipped words and specifically marked dysfluencies, partially pronounced words and laughter. A detailed description of our segmentation and transcription guidelines can be found in [6]. We have found that the transcription word error rate at this stage had been significantly reduced to about 2% from the original error rate of approximately 8%.

The new transcriptions differ significantly from previous transcriptions in the detail used to mark word types. The new transcriptions differentiate between, coinages, mispronunciations, vocalized noise, partial words and words spoken while laughing. We expect this will provide an opportunity to model each of these commonly occurring problems in conversational speech.

3. QUALITY CONTROL

We have taken several steps to ensure that our released data is of the highest possible quality. After our conversations have been validated, we run a set of scripts on the transcriptions which check for different kinds of problems. The final quality of transcriptions generated is verified by regular cross-validation experiments.

3.1. General Checks

First, we use a script which verifies that each word in the transcriptions is also present in the SWB dictionary. Dictionaries related to SWB have evolved over the years and have been developed at several sites. This has resulted in many unseen words being present in the dictionaries. Doing the above check

guarantees that the transcriptions and the lexicon have one-to-one mapping.

The next quality check uses a script, to determine the length of silence-only utterances in the transcription files, flagging those that are less than one second long — our criterion for minimum silence length. Finally, we run a script to ensure that the start time of every utterance is equal to the end time of the previous utterance. This script also makes sure that the end time of the last utterance is equal to the size of the file up to six significant digits.

3.2. Quality of Transcriptions

To gauge the quality of the transcribers we perform regular cross-validation experiments. This ensures that the transcribers maintain a high standard of work. In general, each transcriber validates the same conversation and their transcriptions are compared for accuracy and consistency. We also compare the original LDC transcriptions to the reference, to provide an estimate of the improvement in SWB transcriptions after resegmentation. The current average WER is 2.68% as compared to 8% in the original LDC transcriptions. Table 1 shows the error modalities that remain in the final transcriptions.

Error Modality	% of total errors
transcription of contraction as two words	36%
deletion	27%
insertion	6%
substitution	31%

Table 1: Error modalities in the final transcriptions

4. EXPERIMENTS AND RESULTS

We have completed resegmentation and transcription corrections of 525 conversations of the SWB training corpus. A summary of the modalities of this data is displayed in Table 2. The corresponding numbers on a similar subset from the WS'97 data [7] are also provided. We have observed that of the 100 most frequent words in the resegmented database, 69 are monosyllabic and account for 53% of the total transcription. In the WS'97 data monosyllabic words constitute 74 of the top 100 words and cover 67% of the transcriptions.

4.1. Acoustic Model Adaptation

To estimate the impact of the resegmented training data on recognition performance, we needed to train new acoustic models. We decided to adapt existing acoustic models to this data and evaluate on existing lattices as this was a faster way of getting a preview of the potential improvements in WER.

A word-internal triphone system [7] was used to bootstrap the seed models. The training set consisted of 376 conversations (about 20 hours of speech including silence, or approx. 27500 utterances) common to the baseline training. Four passes of reestimation were carried out. Since the baseline system lacks a laughter model, laughter was used to update the silence model; while words containing laughter were substituted with their baseform.

4.2. Lattice Rescoring

These word-internal acoustic models were used to rescore WS'97 dev test set lattices (the transcriptions of which have already been corrected as described in [7]). The performance of the adapted models (see Table 3) shows a 1.9% absolute improvement over the baseline system. It also reduces the error rate due

Property	ISIP	WS'97
Total duration	71.03 hrs	14.67 hrs
Number of utterances	50334	27717
Average utterance duration	5.08 sec.	1.91 sec.
Number of words	419756	161762
Average words/utterance	13	6
Number of silence utterances	17119	—
Average silence duration	5.53 sec.	—
Occurrence of dysfluencies	20318	7958
Occurrence of laughter-words	1955	—
Number of partial words	2808	—

Table 2: Comparative statistics of ISIP's resegmented SWB transcriptions with those used for WS'97. The above numbers are for the same 525 conversation subset.

Error Rate	ISIP	WS '97
word error rate	**47.9%**	**49.8%**
correct words	55.8%	53.1%
substitutions	31.6%	32.2%
deletions	12.6%	14.8%
insertions	3.7%	2.9%

Table 3: Recognition performance with acoustic models adapted from the resegmented training data. The WER is better than the baseline by almost 2%.

to substitutions and deletions, the main contributors to the error rate on SWB evaluations.

Of the total errors, 63.3% are related to monosyllabic words and 4.7% are due to the various dysfluencies. This is significantly lower than the baseline system which had more than 70% of the errors due to monosyllabic words [7].

5. CONCLUSIONS

We have created a subset of the SWB training corpus that is much more accurate and complete in terms of accounting for silences and partial/laughter words etc. The policy of segmenting at long pauses and *natural* boundaries has allowed the new transcriptions to contain ample acoustic and linguistic context for improved recognition.

We have also demonstrated the potential benefits of using a cleaner database for training of good quality acoustic models. It should be noted that we have achieved a significantly better performance (an improvement of 1.9% absolute) simply by adapting existing models to a small subset of the training data. We believe that building explicit models for laughter and other dysfluencies, as well as for partially pronounced words will result in an even more improved set of acoustic models.

Also, the more accurate transcriptions can be used to adapt LMs that closely reflect the modalities of conversational speech, and reduce the LM mismatch which contributes heavily to the poor recognition. We expect to see a significant improvement in the performance on SWB as a result of training detailed models for highly frequent multi-word contexts. This will also serve as a springboard for studying more complex cross-word phenomena in SWB.

It is evident that proper segmentation and quality transcriptions of training data are essential to achieve better acoustic modeling. We intend to complete the resegmentation of the entire SWB corpus in the near future, thus making a database of much better quality available to the speech research community. We have placed all data, software tools and models developed as part of this work in the public domain [6].

6. REFERENCES

1. J. Godfrey, et al, "SWITCHBOARD: Telephone Speech Corpus for Research and Development", in *Proc. of the IEEE ICASSP*, vol. 1, pp. 517-520, San Francisco, CA, USA, March 1992.

2. A. Ganapathiraju, et al, "Syllable — A Promising Recognition Unit for LVCSR", in *Proc. of the IEEE ASRU Workshop*, pp. 207-214, Santa Barbara, CA, USA, December 1997.

3. B. Wheatley, et al, "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech", in *Proc. of the IEEE ICASSP*, vol. 1, pp. 533-536, San Francisco, CA, USA, March 1992.

4. S. Greenberg, "The Switchboard Transcription Project," *Technical Report of the 1996 LVCSR Summer Research Workshop*, Johns Hopkins University, Baltimore, MD, USA, April 1997.

5. A. Ganapathiraju and J. Picone, "Echo Cancellation For Evaluating Speaker Identification Technology," in *Proc. of IEEE Southeastcon*, pp. 100-102, Blacksburg, Virginia, USA, April 1997.

6. J. Shaffer, et al, "SWITCHBOARD," *http://www.isip.msstate.edu/resources/technology/projects/1998/switchboard*, Institute for Signal and Information Processing, Mississippi State University, 1998.

7. G. Doddington, et al, "WS'97 Syllable Team Final Report," *Tech. Report of the 1997 LVCSR Summer Research Workshop*, Johns Hopkins University, Baltimore, MD, USA, December 1997.