RESEGMENTATION OF SWITCHBOARD

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, Joseph Picone

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

ABSTRACT

The SWITCHBOARD (SWB) corpus is one of the most important benchmarks for recognition tasks involving 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 is imperative to improve the quality of segmentations and transcriptions of the training data to achieve better acoustic modeling. By adapting existing acoustic models to only a small subset of such improved transcriptions, we have achieved a 2% absolute improvement in performance.

1. INTRODUCTION

One of the most challenging tasks for current state-of-the-art LVCSR systems is to accurately recognize telephone conversations. The SWB corpus [1] is currently the standard benchmark for such applications. It contains 2430 conversations averaging 6 minutes in length; in other words, 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.

This is a very challenging task notwithstanding the limitations posed by the telephone channel, including bandwidth, transducer, noise and echo. Fast speaking rates; poor coarticulation at word boundaries; 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. Moreover, 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. The result is poor acoustic modeling for recognition, and a high degree of mismatch between the training and test data.

One approach to reduce this acoustic-level mismatch is to predict a large number of common alternate pronunciations and incorporate these into the acoustic models as additional paths. Such pronunciation modeling techniques suffer from related problems of intelligent integration of language model and acoustic model scores, and have met with limited success. Instead, the acoustic models can be reestimated to automatically incorporate such pronunciation variations. This requires high quality transcriptions and segmentation of the training database.

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. Segmentation is typically automatic and uses techniques based on energy levels, information-based metrics and phone-level recognition [3]. However, these introduce unnatural breakpoints in the utterances, thus decreasing the

effectiveness of the LM. On the other hand, linguistically motivated segmentation [4] often results in extremely short phrases that do not provide sufficient acoustic context for accurate recognition.

We have sought to balance this trade-off by manually resegmenting the automatic segmentations and ensuring ample context for both acoustic and language modeling applications. Our approach to resegment the data consists of:

- echo cancellation
- 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

2.1. Echo Cancellation

Echo is a major cause of transcription errors on SWB. Besides interference, it often causes wrong channel assignments on the original data; because it is very hard to identify which channel corresponds to which speaker. We remove echo with our standard least mean-squared error-based echo canceller [5].

2.2. Utterance Resegmentation and Correction

We discovered that a vast majority of the problems with the current SWB segmentations is due to segment boundaries being placed between words which have little or no acoustic separation. These segments when split between words with a high degree of coarticulation have an adverse effect on the training of models.

We believe that utterances delimited by sufficiently lengthy pauses or *natural boundaries* such as sentence/phrase ends or speaking turns can maintain both acoustic continuity as well as linguistic context. Therefore our strategy for resegmentation is:

- merge utterances which are currently split at counter-intuitive points
- segment at locations where there is clear silence separating each segment
- segment along phrase, sentence, and/or train-of-thought boundaries

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 (about 0.4-0.5 sec of silence at each end) and/or turn boundaries.

Besides adjusting the utterance segment boundaries, we also corrected 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 10%.

2.3. Word Alignment Review

The new segmentations and transcriptions of the training data were used to create a new set of word alignments by performing supervised training with our best phone-based recognizer. These word alignments were reviewed manually to further improve the accuracy of the transcriptions. This step caught most of the errors that had slipped through the transcription process, thus resulting in an extremely accurate database (a cross-validation test on sample utterances places the WER at 1%).

3. THE SEGMENTER

We developed a graphical, point-and-click interface tool to expedite the segmentation/transcription process. This tool, written in a mixture of C++ and Tcl/Tk, is designed to be highly portable across platforms (we currently run it on Sun Sparcstations as well as Pentium-based desktops running Solaris, an extension to Windows is currently in preparation). It also supports numerous audio utilities. The current version of the segmenter is highly customized to be used with the SWB corpus. However, it is easily extensible to other domains and is freely available.

The segmenter supports both mono and stereo audio. Stereo audio is an integral part of the SWB task, since it allows the transcribers to probe each side of the conversation separately or listen to the full context. This, coupled with the echo cancellation of

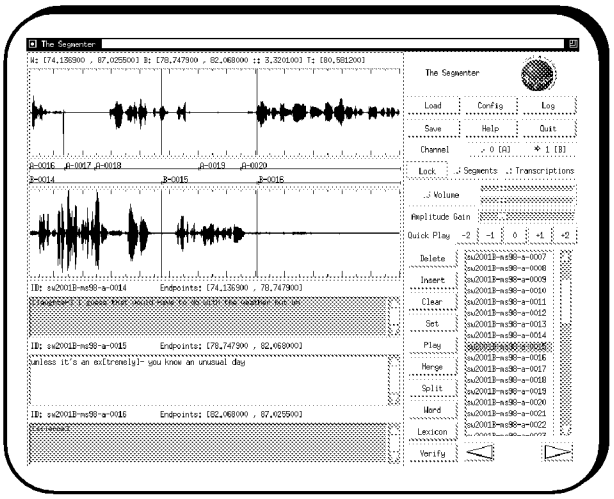

Figure 1. A screen-shot of the Segmenter

data, allows them to fix many of the swapped channel problems that have plagued SWB. The display area of the tool provides instant access to the acoustic waveforms as well as the audio context for any utterance, plus the functionality to zoom in and/or play a selected portion of the utterance. The word alignment mode allows checking the transcription word-by-word, thus providing a quick but efficient means of strict quality control at a manageable cost.

The segmenter has allowed transcribers to achieve a throughput of less than 20x real-time on SWB (correction of existing transcriptions for a two-sided 5 minute conversation typically requires slightly less than 100 minutes). Figure 1 shows a sample screen-shot of the segmenter interface.

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 1. The corresponding numbers on a similar subset from the WS'97 data [7] are also provided. It can be seen that our transcriptions and segments are significantly more detailed, with explicitly marked silence, dysfluencies, laughter, and partial words. The WS'97 set had these omitted, besides partially skipping some conversations. We also 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.

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

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 2) shows a 1.9% absolute improvement over the baseline system. It also reduces the error rate on substitutions and deletions, the main contributors to the error rate on SWB evaluations.

Of the total errors, 63.3% are attributable to monosyllabic words and 4.7% are due to the various

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

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, E. Holliman and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development", in *Proceedings of the IEEE ICASSP*, vol. 1, pp. 517-520, San Francisco, CA, USA, March 1992.

2. A. Ganapathiraju, V. Goel, J. Picone, A. Corrada, G. Doddington, K. Kirchoff, M. Ordowski, and B. Wheatley, "Syllable — A Promising Recognition Unit for LVCSR", in *Proceedings of the IEEE ASRU Workshop*, pp. 207-214, Santa Barbara, CA, USA, December 1997.

3. B. Wheatley, G. Doddington, C. Hemphill, J. Godfrey, E. Holliman, J. McDaniel and D. Fisher, "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech", in *Proceedings 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 *Proceedings 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. A. Ganapathiraju, et al, "WS'97 Syllable Team Final Report," *Technical Report of the 1997 LVCSR Summer Research Workshop*, Johns Hopkins University, Baltimore, MD, USA, December 1997.