SELECTION FOR ACOUSTIC COVERAGE FROM UNLIMITED SPEECH EXTRACTED FROM CLOSED-CAPTIONED TV

Photina Jaeyun Jang, Alexander G. Hauptmann
Computer Science Dept, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, Pennsylvania 15213
{jaeyun, alex}@cs.cmu.edu
http://www.cs.cmu.edu/~{jaeyun,alex}

ABSTRACT

Given unlimited amounts of speech training data, it is desirable to predict informative subsets that will still improve the resulting acoustic model. We present a triphone frequency threshold measure for predicting informative subsets from vast amounts of speech. Results with single pass decoding show that acoustic models built from our selection-based speech set perform better than when trained on similar amounts of non-selected speech, and perform similar to models built from the original, larger amount of speech.

1. INTRODUCTION

Large amounts of hand-transcribed speech training data are required to build or improve speech recognition systems using current technologies. Speech data collection and its expensive manual transcription have been a bottleneck for improving speech systems. LDC (Linguistic Data Consortium) has made available 200 hours of manually transcribed speech in 1998 and about 600 hours will be available in 1999. Additionally, we have demonstrated a technique for automatically extracting speech data together with automatically generated accurate transcriptions from daily broadcast television stored as MPEG video [4]. Speech recognition research is moving towards training from practically unlimited quantities of transcribed speech, with the associated increase on the cost of the training process. If one could predict the critical subsets of speech data that add significantly to the performance of a speech recognition system, then it would be possible to avoid training acoustic models on the complete speech set, much of which is very redundant, thus reducing processing effort.

We will first give an overview of the work on automatic extraction and transcription of unlimited amounts of speech and then describe the predictive selection of informative subsets from this large pool of speech.

2. VIRTUALLY UNLIMITED AMOUNT OF SPEECH

A method for solving both the problems of expensive speech data collection and of expensive human annotation of speech was introduced in [4]. The Informedia Digital Video Library [1, 3] records and digitally stores television broadcasts in MPEG format. If there is closed captioning available, the closed captioned data is stored separately as text.

2.1 Extracting speech from MPEG video.

The MPEG audio stream is split from the MPEG video and is uncompressed into its original 44.1 kHz 16bit sampling rate and then downsampled to 16 kHz. The audio is further processed into Mel-frequency coefficients to represent the audio signal as feature vectors of 12 values every 10 milliseconds.

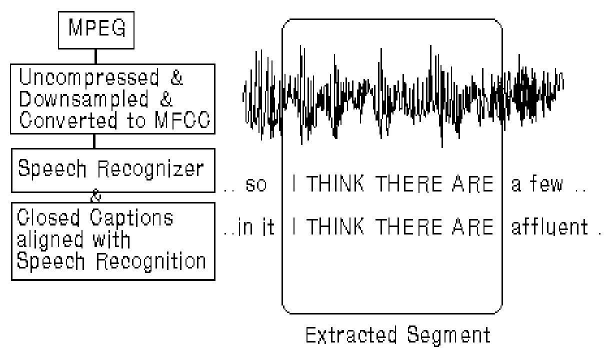

Figure 1: Extracting an audio *segment* with a reliable transcription from the MPEG video using the closed captions aligned with a rough speech recognition transcript.

2.2 Generating Accurate Transcriptions

Two errorful sources of transcriptions are available for the MPEG audio: the broadcast closed-captions and the initial output of the Sphinx speech recognizer [7].

To obtain more accurate transcription of an audio stream, the closed-captions and the decodings of the speech recognizer are aligned. Sequences of words in the closed captions and the recognition output are matched with a dynamic programming alignment [6]. From the alignment, segments are selected where sequences of three or more words are identical in both the closed captions and speech decodings. Figure1 shows a selection of the words "I think there are" from the audio stream, using the associated speech recognition and closed captions.

The effect of this selection is to verify the closed captions using an independent source of the speech recognition output. We view the matching word sequence of closed-captions and the speech recognizer's hypothesis as a form of mutual confirmation, or as a binary confidence measure. The corresponding audio segment for the selected annotation segment is extracted and added to the speech training set. Since the errors made by the human captioning service and those made by the Sphinx speech recognizer are largely independent, extended sections over which the captions and the Sphinx transcript correspond have been (mostly) correctly transcribed.

The amount of transcribed speech data is thus virtually unlimited, depending only on the amount of available television broadcasts with associated closed-captions.

3. ACOUSTIC MODELS BUILT FROM LARGE AMOUNTS OF UNSELECTED SPEECH

111.5 hours of training data obtained this our automatic extraction and transcription method improved the best acoustic model for Sphinx-III built from the 1997 DARPA HUB4 broadcast news model (62.8 hours) by 7.58% relative (2.49% absolute). Thus the CCtrain models used a total of 174.3 hours of training data. For all three different language model weights, the CCtrain acoustic model achieved improvements over the CMU HUB4'97 baseline system on the DARPA HUB4 Dev'96 test set.

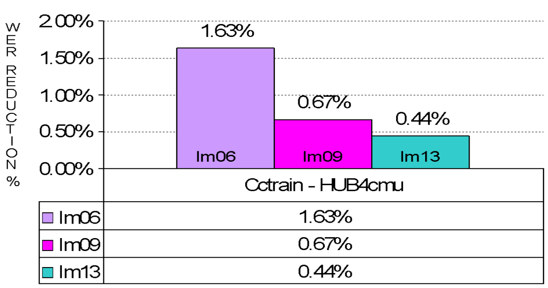

Figure 2: The absolute Word Error Rate (WER) reduction of the CCtrain system over the HUB4cmu baseline system, single pass decoded with language model weights 6, 9 and 13.

These results show that higher language model weights make it harder to estimate the degree of improvement on the acoustic model. This is due to the increasing contribution of the language model to the speech recognition accuracy during the decoding process.

For all subsequent figures, the language model weight was set to 6 to demonstrate the effect of the acoustic component, with relatively little influence provided by the language model.

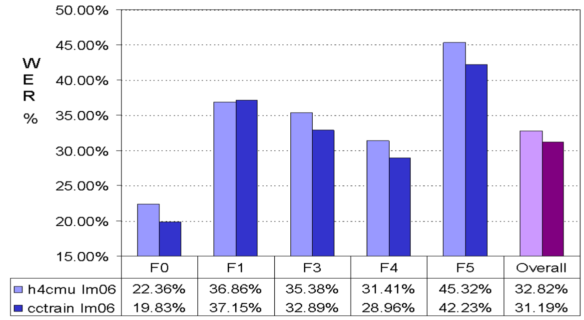

Figure 3: WER on the different HUB4 acoustic conditions with single pass decoding. The last two columns show the overall WER.

Figure 3 shows the Word Error Rate (WER) for the five different acoustic conditions, contrasting the baseline CMU HUB4 system (each 1st column) with the adapted CCtrain system (each 2nd column). The CCtrain model provides superior results on the F0 (baseline broadcast speech), F3 (speech with background music), F4 (speech with degraded acoustic conditions) and F5 (speech from non-native speakers). For the F1 (spontaneous broadcast speech) condition the performance is similar.

4. ERROR ANALYSIS

The word errors on the Dev'96 test set presented above, were examined with respect to the number of training instances in the HUB4 and the CCtrain set. Only words that were present in both the training corpus and the test set were examined. The x -axis represents the number of times a word was present in the training set and the y-axis is the number of errors of that same word in the test set.

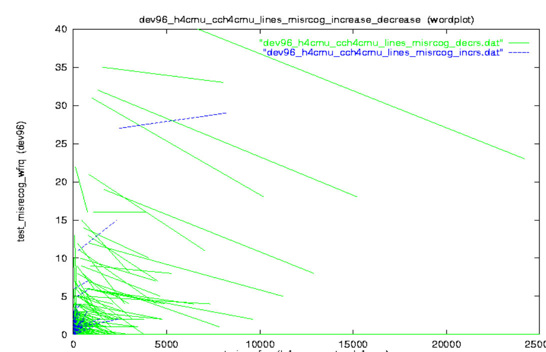

Figure4: Effect of word frequency in the training corpus (x-axis) on the number of errors in the test set (y-axis). A descending line reflects an improvement in word error by the CCtrain model (82.6% of the test set words), and an ascending line shows an increase in word error (only 17.4% of the test set words).

Each line on Figure 4 represents a word in the test set, each connected with a HUB4 point and a CCtrain point. A HUB4 point represents the number of training instances for a word in the HUB4'97 training corpus and the number of errors for this word of the HUB4 baseline system, and vice versa with the CCtrain point. Since the CCtrain corpus was comprised of the HUB4'97 training set as well as our automatically extracted new training material, the CCtrain point is always on the right end of each line, while the HUB4 point is on the left end of each line. 82.6% of the words showed improvement in recognition error, and only 17.4% of the words increased in error, decoded with the CCtrain model.

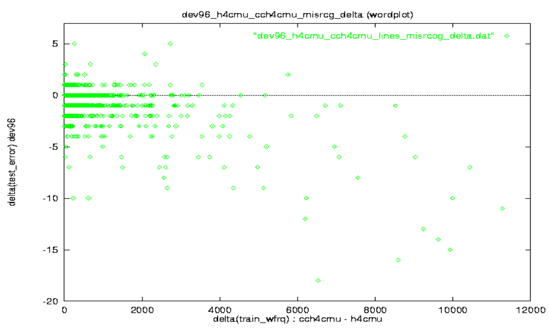

Figure 5: A delta plot for the difference in errors for each word of both the CCtrain and HUB4 models on the test set (y-axis), and the training word frequency (x-axis). WER reduction of CCtrain are the negative values (82.6% of words), whereas the increase of WER are positive values on the y-axis (only 17.4% of all words).

The same statistics is plotted in Figure 5. The CCtrain word errors are subtracted from the HUB4cmu errors, and the difference in training frequencies between CCtrain and HUB4cmu are plotted on the x-axis. Each data point in Figure 5 represents the increase in word training frequency for the CCtrain corpus on the x-axis and the difference in word errors on the y-axis between the two conditions. If the word error does not change, the difference in errors is 0. If the HUB4cmu word errors were lower than the CCtrain error rate, then the data values will lie above the 0 point on the y-axis. If the CCtrain showed an improvement for that word, the point will be plotted below 0 on the y-axis.

5. PREDICTING INFORMATIVE SUBSETS OF TRAINING SET

For a more efficient use of the potentially unlimited amount of automatically derived speech training data, we want to extract only those subsets, which provide new information to improve the acoustic model.

The automatically obtained speech transcripts are converted to triphone sequences by mapping the lexical words into the corresponding set of triphones. A triphone is a context dependent phoneme with left and right phoneme context information.

5.1 Error Analysis on Triphone Frequency

If there is little or no noise in the corpora to train the acoustic model, the more often triphones or words are observed in the training set, the more examples are learned. Therefore we would expect better acoustic models. This is indeed the case as shown in Figure 6, where the triphone error rate levels off at about 1000 instances of triphones in the training set.

The triphone frequency in the training set and its corresponding error rate in the test set is plotted on the x- and y-axis respectively. The `trend' line shows the averaged error rate for triphones within a histogram bin of 30 similarly frequent words. From this plot we observed that the test error rate decreases asymptotically as the triphone frequency of the training instances increases. As more triphones are observed in the training set, the test error rate on those triphones decrease. After about 1000 triphone training instances, only a minimal decrease in error rate was observed.

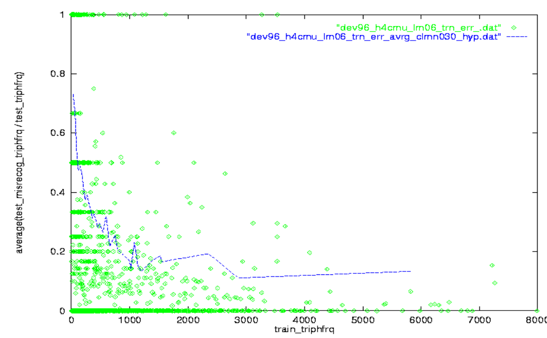

Figure 6: The triphone error rate (y-axis) for each triphone in the Dev'96 test set, with the corresponding triphone frequency in the training set (x-axis).

A similar phenomenon was observed when plotting the word frequency in the training set (x-axis) and the corresponding word error rate in the test set (y-axis). To show the trend of the individual points, a line was again plotted that represents the average word error rate in histogram bins, where each bin combines 30 adjacent frequencies.

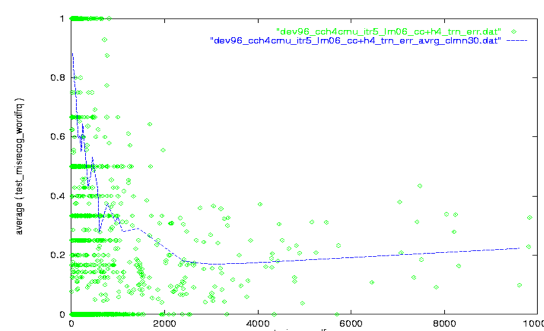

Figure 7: WER (y-axis) for each word in the Dev'96 test set, with corresponding number of word frequency in the training set (x-axis).

The analysis of the Dev'96 test set indicated that training on about 1000 instances of a triphone or 2000 instances of each word would be sufficient to obtain acceptable recognition performance. To verify this hypothesis, we conducted an experiment with selective training by excluding redundant training data, which had no instances of triphones below the threshold.

6. PERFORMANCE ON TRIPHONE FREQUENCY THRESHOLDED, SELECTED SPEECH

We hypothesized a threshold, set to 1000, on the triphone frequency where the triphone error rate started to closely approximate the asymptotic upper bound. A subset of our automatically transcribed speech training data was extracted by adding a new segment to the subset only if the frequency of any triphone within the segment was below our hypothesized threshold. We trained an acoustic model on this training subset to see whether the hypothesized threshold near the asymptote on our plot is a good indicator for the number of triphone observations required.

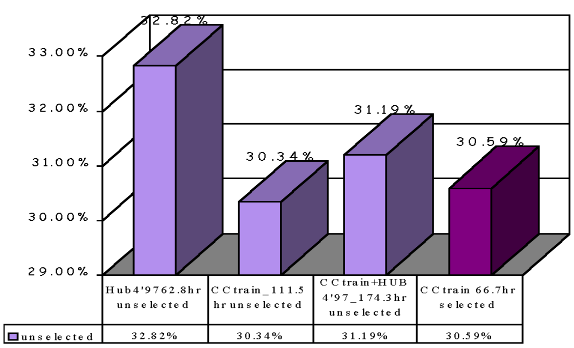

Figure 8: WER comparison of models built from three non-selected sets (62.8 hours HUB4 baseline, 111.5 hours CCtrain, 174.3 hours combined) and one selected set (66.7 hours selected from 111.5 hours of Cctrain data).

The results with single pass decodings are shown in Figure 8. Our selected training subset of 66.7 hours, which included only approximately 60% of the original training set comprised of 111.5 hours, yielded a word error rate improvement of 6.76% relative (2.22% absolute) over the Sphinx baseline system built from 62.8 hours of HUB4'97 training data. However, our selection technique resulted in a WER increase of 0.25% compared to the complete 111.5 hours of CCtrain data set.

The 111.5 hours of unselected CCtrain data lead to a 0.8% lower WER than when combined with the HUB4'97 yielding to 174.hours. We believe this is due to the fact that our automatic extraction technique produces more accurate transcriptions than those found in the manual HUB4'97 training data.

It may appear contradictive to first extract an arbitrarily large amount of transcribed speech and then to reduce that set into a smaller subset. But selecting a subset from the original, larger amount of speech, increases the probability to incorporate more diverse triphones.

7. CONCLUSION

Selection from collected speech data based on triphone frequency thresholding reduces training costs, results in better decoding performance than when similar amount of unselected manually transcribed speech is used, and performs comparable to much larger amounts of speech training sets.

For a more powerful measure to predict subsets leading to better performance, than when utilizing the original larger set, we should consider that certain triphones may require more diverse acoustic examples, that some triphones may need to be redefined into more specific or general units, as well as the quality of the audio signal for the particular triphone examples, that acoustically noisy instances need to be excluded or compensated with more training instances, etc.

However, acoustic models built from selected training sets based on triphone frequency thresholding, as introduced in this paper, perform better than those that were trained on similar amounts of unselected, manually transcribed speech.

8. ACKNOWLEDGEMENT

This paper is based on work supported by the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA) and the National Aeronautics and Space Administration (NASA) under NSF Cooperative agreement No. IRI-9411299

9. REFERENCES

[1] Christel M., Kanade T., Mauldin M., Reddy R., Stevens S., and Wactlar H. (1996), "Techniques for the Creation and Exploration of Digital Video Libraries", *Multimedia Tools and Applications*, *Kluwer Academic Publishers*.

[2] Hauptmann, A.G. and Wactlar, H.D. (1997), "Indexing and Search of Multimodal Information", *International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*

[3] Informedia CMU,http://www.informedia.cs.cmu.edu

[4] Jang, Ph.J., Hauptmann, A.G. (1999), "Improving Acoustic Models with Multimedia Audio Speech", *IEEE International Conference on Multimedia Computing and Systems (ICMCS'99)*.

[5] LDC, http://morph.ldc.upenn.edu/ldc/frame.html

[6] Nye, H. (1984) "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition." *IEEE Transactions on Acoustics, Speech, and Signal* Processing.

[7] Sphinx CMU, http://www.speech.cs.cmu.edu