AUTOMATIC TRANSCRIPTION OF COURTROOM SPEECH

Rohit Prasad, Long Nguyen, Richard Schwartz, John Makhoul
BBN Technologies
10 Moulton Street, Cambridge, MA 02138, USA

{rprasad, ln, schwartz, makhoul}@bbn.com

ABSTRACT

In this paper we describe our on-going effort in developing a speech recognition system for transcribing courtroom hearings. Court hearings are a rich source of naturally occurring speech data, much of which is in public domain. The presence of multiple microphones coupled with presence of noise and reverberation makes the problem simultaneously rich and challenging. We have exploited the availability of multiple channels to mitigate, to some extent, the noise problem prevalent in courtroom speech. By using a novel technique for channel change detection, domain-specific language modeling, and unsupervised channel adaptation we have been able to achieve a word error rate (WER) of 36% on actual courtroom hearings. We also report on acoustic modeling experiments using “legal” transcripts for 120 hours of court hearings in a lightly supervised mode.

1. INTRODUCTION

The benefits of having a system that can automatically record, transcribe, index, and archive minutes of court hearings are many and obvious. Court hearings challenge existing speech and language technologies in a number of dimensions, including the presence of various types of noise, reverberation, simultaneous speakers, and multiple distant microphones. In this paper we describe our exploratory work in developing a speech recognition system for transcribing courtroom hearings.

To develop an understanding for the problems posed by courtroom speech, we started acquiring a corpus of real-world court hearings. In particular, we have partnered with FTR Pty Ltd., a leading supplier of court and hearing room digital recording systems, to supply us with digital speech data and the corresponding “legal” transcripts. Till date, we have received 165 hours of audio data for 125 courtroom sessions.

In Section 2, we describe the audio data collection process in the courtroom and the various characteristics of the data. In Section 3, we present our initial recognition results with a small amount of supervised acoustic training and domain-specific language modeling. Section 4 describes a number of recognition accuracy improvements with acoustic models trained on broadcast news data, which have resulted in a word error rate (WER) of 36% on this difficult task. In Section 5, we report recognition results with acoustic models trained on a large amount of courtroom speech in a lightly supervised mode using the approximate legal transcripts.

2. COURTROOM CORPUS

2.1. Courtroom Speech Data

When recording a courtroom proceeding digitally, typically multiple microphones are placed in different positions in the courtroom: one for the judge, one for the witness, one on each of the two attorneys’ desks, and the others are located next to the jury and other locations. It is not unusual to have as many as eight microphones in a courtroom. The digital audio signals from the different microphones are then mixed and recorded over either 2 or 4 channels. A person in the courtroom monitors the recording and enters log notes, which include timestamps and the names of the people speaking in the courtroom. Later, a human transcriber prepares a transcript of the session by listening to the recording and typing the transcript into a computer.

The data we have received from FTR includes the audio data, the log notes, as well as the legal transcript. Most of the audio files are 4-channel audio with a sampling frequency of 22.05 KHz. The files are time-stamped and are of 5 min. duration each. Our study has indicated that the quality of the audio data is highly variable. The microphones around the courtroom pick up various types of background noise and non-speech events like coughing, laughing, whispering, paper rustling, etc. We have also observed that the speaking style in the courtroom is very spontaneous. This is illustrated in the following passage from a judge.

I also appreciate uh Mr. uh, uh, Jessup’s um uh willingness to go forward and uh be uh as forthcoming as uh he uh can. But I, I, I still think we’re on a little bit of a precipice here, Mr. Kennedy.

Besides the quality of speech mentioned above, the legal transcripts presented a number of problems for our work. One problem was the lack of timestamps in the transcripts, so it is not easy to match the transcripts to the speech. Another problem is that some audio segments were never transcribed, while in other cases there was text with no corresponding audio. Moreover, these transcripts were generated to provide content rather than providing faithful annotation for training acoustic models. So, they lack most of the non-speech events and do not include hesitations and restarts.

2.2. Test and Training data

The criterion we used for selecting the test data was the session duration. We chose the 8 smallest sessions as our test set, which totaled 6 hours of audio. The remainder of the data we planned to
use for acoustic and language training. On the average there are 5 speakers per test session and tend to be primarily male speakers. The test sessions covered topics such as, corporate bankruptcy, insurance coverage, driving under influence, and car accident.

Since the legal transcripts are not accurate enough for evaluation purposes, we decided to have the entire test set transcribed, i.e., each of the 5 min. audio chunks had an accurate transcription associated with it. For controlled acoustic modeling experiments, we also had 8 hours from 5 sessions transcribed accurately. We found that the mismatch, in terms of WER, between accurate transcripts (AT) and the legal transcripts (LT) from FTR was 25%, as illustrated in the alignment result below. This means that we can not use the legal transcripts as-is for acoustic modeling.

AT: I I I appreciate uh very much um
LT: * * I appreciate ** very much **

3. INITIAL EXPERIMENTS

First for both 8 hours training and 6 hours test data, we separated the multiple data streams from the FTR audio. Then, for convenience, we added the different streams to form a single signal. The resulting 5-min. audio files were segmented into smaller chunks by using Broadcast News (BN) [1] acoustic models for Viterbi alignment of the transcript with the audio. And then, segmenting during silence segments of at least 200 ms.

We trained speaker and gender independent acoustic models on 8 hours of manually transcribed data. To enhance robustness to channel variations, the features extracted from the speech signal were normalized using RASTA [2] processing. We estimated triphone models using Phonetically-Tied Mixtures (PTM) and State-Clustered Tied Mixtures (SCTM). We also trained both bigram and trigram language model with a simplified backoff strategy [3] on legal transcripts from 28 FTR sessions having 400K words and 22M words of text from court proceedings downloaded from the Web. The lexicon size for the language model was 52K words, and the out-of-vocabulary rate was 1.1%.

The recognition system used was a simple configuration of the BBN BYBLOS system [1], which employed the two-pass N-best decoder [4]. The first pass uses a single-phonetic-tree fast-match algorithm with PTM models and a bigram grammar [5] to constrain the acoustic search space. In the second pass, the decoder uses the SCTM non-crossword models and a trigram grammar to generate the N-best hypotheses. The N-best hypotheses are rescored using SCTM crossword models and reordered to output the top-1 hypothesis as the final recognition result.

We decoded the entire 6 hours test set with the acoustic models trained on 8 hours of domain-specific data and also with an acoustic models trained on 150 hours of BN data [1]. The BN acoustic models employed quinphones, instead of just triphones. Both decodings used the language model mentioned above. The resulting WER with acoustic models trained on a small amount of court hearings was 61.4% as shown in Table 1. The WER with BN acoustic models was 58.4%.

The high WER of these initial recognition experiments clearly shows that transcription of courtroom speech is a challenging task. An obvious strategy for reducing the WER on this task would be to train acoustic models on more than 100 hours of courtroom speech from different hearings. But, as mentioned earlier the legal transcriptions made available to us from FTR can not be used for acoustic training as-is. Therefore, we decided to first focus on exploiting certain characteristics of the courtroom data to significantly reduce the WER using the BN acoustic models, and then develop strategies for utilizing the legal transcripts to train better acoustic models on the in-domain data.

4. ACCURACY IMPROVEMENTS

First, we split the 6 hours test set into 2 sets of approximately the same duration. One set we plan to use as a development set, and the other as a validation set. We report recognition improvements on the 3 hour development set using the BN acoustic models.

4.1. Channel Change Detection

In the experiments described in Section 3, for convenience we had summed the channels to form a single signal. It became clear by subjective listening, that summing the channels had introduced significant distortion in the speech signal. We observed that in the courtroom corpus, most of the time, the speaker is dominant on a single channel. Rest of the channels for a speaker turn do pick up some speech, but are primarily a source of noise and other non-speech events. Therefore, we decided to investigate recognition on individual channels for improving the WER.

Under the assumption that speakers are dominant on a single channel, we developed a channel change detection procedure. First, we detect speaker changes on the summed channel signal. Next, for each speaker turn we select the channel with the highest signal-to-noise ratio (SNR). We define SNR(DB) as the difference between the 80th and 20th percentile log energy over the entire speaker turn. For speaker change detection, we used the novel 2-stage approach developed for transcribing broadcast news [6, 7]. In this approach, the first stage detects speech/non-speech boundaries using a coarse gender-independent phoneme recognition. We used an 8-class phone model trained on 20 hours of broadcast news: three broad phonetic classes (vowels, fricatives, and obstruents) and five models for non-speech phenomena (silence/noise, laughter, breath, lip-smack, and music). The second stage performs the actual speaker segmentation by hypothesizing a speaker change boundary at every phone boundary that was located in the first stage using a penalized likelihood ratio criterion.

The resulting WER when decoding the highest-SNR channel for each speaker turn was 43.0% as shown in Table 2. The corresponding WER for decoding on the summed channel was 59.6%. Therefore, by detecting channel changes and performing recognition on individual channels, we have reduced the WER by more than 16% absolute.

<table>
<thead>
<tr>
<th>Acoustic Source</th>
<th>Amount</th>
<th>#Gaussians</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Court Hearings</td>
<td>8 hrs</td>
<td>32K</td>
<td>61.4%</td>
</tr>
<tr>
<td>Broadcast News</td>
<td>150 hrs</td>
<td>277K</td>
<td>58.4%</td>
</tr>
</tbody>
</table>

Table 1: Initial recognition results using domain-specific and broadcast news models
We investigated the quality of SNR as a criterion for selecting the best channel. To get the lower bound on the WER for channel selection, we decoded all 4 individual channels for each speaker turn, and then selected the hypothesis with the least WER. The overall WER was 42.4%. Assuming there were no errors in the detection of the speaker changes, this 0.6% absolute difference clearly shows that SNR is a good criterion for selecting the best channel.

4.2. Improved Language Model

As mentioned in Section 3, we had trained a language model on legal transcripts from 28 FTR sessions and 22M words of text downloaded from the Web. Both these sources do not correspond to the actual acoustics of the court proceedings. Therefore, we added the text from the 8 hours accurately transcribed data to our language modeling data. This particular data set had a faithful annotation of phenomenas like false starts, pause fillers, etc. We also added legal transcripts for another 27 FTR sessions that were made available to us recently. Decoding the same segments as in Section 4.1, with BN acoustic models and the new language model reduced the WER to 42.1%.

4.3. Unsupervised Channel Adaptation

The channel characteristics for courtroom speech is significantly different from broadcast news data. The recording conditions may vary from one session to the other. Moreover, each hearing has a few dominant speakers, primarily the attorneys who speak for a long duration. Therefore, adapting the BN speaker independent (SI) acoustic models to courtroom speakers or channels seems very appropriate in this cross-domain recognition task.

For adapting to individual channels, we first decode the highest-SNR channel for each speaker turn as in Section 4.2. We then group all the audio segments and corresponding recognition hypotheses for each of the individual channels. Finally, we use the recognition hypotheses to adapt the mixture densities’ parameters using maximum likelihood linear regression (MLLR) [8].

We decoded the same test set with channel-adapted mixture densities and the improved language model. The resulting WER was 36.2% as shown in Table 2 - a 14% relative reduction in comparison to using BN-SI acoustic models.

We also analyzed the characteristics of the test sessions in detail. In Table 3, we present features such as, number of speakers, number of words, perplexity, SNR, and WER for each session. Our study shows that there is very little correlation between any of the above mentioned features and the WER for the session.

5. IN-DOMAIN ACOUSTIC MODELING

Since supervised training requires an accurate orthographic transcription of the audio, and since obtaining such accurate transcriptions is very costly and time consuming, we decided to investigate methods that could take advantage of the legal transcripts that are being provided to us along with the audio data, similar to those described in [9, 10].

5.1. Training with Individual Channels

The first experiment we performed was to incorporate channel change detection in our acoustic modeling paradigm to train a model on 8 hours of accurately transcribed data. For each of the 5 min. transcript, we derived timestamps for the reference words by Viterbi alignment of the transcript with the summed channel signal. Next, as in Section 4.1, we detect speaker changes on the summed signal and select the highest-SNR channel for each speaker turn. We made a reference for all speaker turns by using the timestamps for the reference words obtained earlier.

We trained speaker and gender independent triphone models with the highest-SNR channel segment and the corresponding transcript. The PTM models used a total of 13K Gaussians, and 9K sets of mixture weights. The SCTM models used about 800 distinct sets of at-most 64 Gaussians for a total of 40K Gaussians, also with a set of 9K mixture weights. Using this acoustic model, we decoded the same test set as in Section 4, with the language model in Section 4.2. After channel adaptation, the resulting WER was 44.4% as shown in Table 4, which is 8.2% absolute worse than using the broadcast news acoustic models.

5.2. Training with Legal Transcripts

In this section we describe our strategy for training acoustic models on 120 hours of raw courtroom data using the legal transcripts and the BN acoustic models as bootstrap models. In [10] it was demonstrated that for broadcast news data the approximate transcriptions in the form of closed-captions can be used for training a powerful language model. Decoding results with this language model changes and using features from individual channels.

We also analyzed the characteristics of the test sessions in detail. In Table 3, we present features such as, number of speakers, number of words, perplexity, SNR, and WER for each session. Our study shows that there is very little correlation between any of the above mentioned features and the WER for the session.
model can be used as transcripts for training acoustic models that result in recognition accuracy comparable to the acoustic models trained with accurate transcripts.

In our strategy, for each session we train a session-specific language model using the legal transcript. This language model is obtained by merging the N-gram counts from the legal transcript for the session with the counts used in training the language model in Section 4.2. Next, we apply the channel change detection procedure on the audio data to select the highest-SNR channel for each speaker turn. The highest-SNR channel for each speaker turn is decoded with BN acoustic models and the session-specific language model. The decoding result is used for unsupervised channel adaptation, and the speech segments are decoded again with adapted parameters. The recognition results from channel adapted decoding are used as transcripts for training acoustic models.

We processed entire 120 hours of raw audio corresponding to 88 sessions according to the procedure described above. Since courtroom speech has significant amount of silence and non-speech events, in reality, we had decoding results for 100 hours of speech segments. We discarded about 25 hours of data by using heuristics such as percentage of noise words and SNR. Therefore, we ended up with 75 hours of speech segments with approximate transcripts.

We estimated triphone models using PTM and quinphone models using SCTM on 75 hours of speech segments resulting from above processing, in addition to 8 hours of accurately transcribed data from Section 5.1. Using this acoustic model, we decoded the test set in Section 4 with the language model in Section 4.2. After channel adaptation, the resulting WER was 38.7% as shown in Table 5. Therefore, by adding 75 hours of automatically processed courtroom speech to 8 hours of accurately transcribed data, we have improved the WER from 44.4% to 38.7%, but the BN acoustic models are still 2.5% absolute better.

![Table 5: Recognition results with acoustic models trained on court hearings using legal transcripts and another set of models trained on union of broadcast news data and court hearings](image)

Since acoustic models trained on 150 hours of accurately transcribed BN data performed better than models trained on 83 hours of courtroom speech, we decided to train acoustic models on the union of these two acoustic sources. We estimated triphone models using PTM and quinphone models using SCTM on 233 hours of audio data. After channel adapted decoding, the resulting WER was 37.1% as shown in Table 5. It seems that adding in-domain data, but with approximate transcripts, into the BN acoustic models does not improve recognition accuracy at all.

### 6. CONCLUSIONS AND FUTURE WORK

We have succeeded in establishing an initial corpus of courtroom hearings. In our exploratory work we have achieved a WER of 36% on actual courtroom hearings, primarily due to channel change detection and channel adaptation. In addition, we have demonstrated that the legal transcripts provided with the audio data can be used to obtain reasonable recognition accuracy. We are still receiving data from FTR which will be used for improving the acoustic models as well as the language model. We will also explore speaker adaptation with speaker adaptive training (SAT). Once the WER is reduced to below 30%, we will focus on metadata extraction for providing a rich transcription.

### Acknowledgments

This work was sponsored by the Defense Advanced Research Projects Agency under Contract No. N66001-99-D-8615 and monitored by SPAWAR. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

The authors would also like to thank FTR Pty Ltd., especially Steve Townsend, for providing the courtroom data and information regarding courtroom proceedings.

### References