Transcript-Dependent Speaker Recognition using Mixer 1 and 2

Fred S. Richardson and Joseph P. Campbell

MIT Lincoln Laboratory
Lexington, Massachusetts, USA

Abstract

Transcript-dependent speaker-recognition experiments are performed with the Mixer 1 and 2 read-transcription corpus using the Lincoln Laboratory speaker recognition system. Our analysis shows how widely speaker-recognition performance can vary on transcript-dependent data compared to conversational data of the same durations, given enrollment data from the same spontaneous conversational speech. A description of the techniques used to deal with the unaudited data in order to create 171 male and 198 female text-dependent experiments from the Mixer 1 and 2 read transcription corpus is given.

1. Introduction

In transcript-dependent speaker recognition, a subject reads aloud from a transcript to produce a speech sample. In law enforcement applications, suspects might be asked to read from a transcript for comparison against speech samples collected as evidence [1]. This process can be used with caution as a comparative test in the forensic domain [2]. A tremendous challenge in these applications is that there is often a mismatch between the speech samples to be compared with respect to variations of the speaker, the speech and voice, and the channel. Here we explore the variability in performance across different transcripts and compare it to that of conversational speech at the same durations.

In text-dependent speaker recognition, the system exploits prior knowledge of the text to be spoken to improve performance. Many applications of text-dependent speaker recognition also benefit from cooperative users and sometimes even matching conditions and matching speech content between the samples. These favorable conditions can be exploited to improve performance of text-dependent speaker recognition of text-independent speaker recognition [3, 4, 5]. In contrast, the approach described in this paper uses a text-independent system which is trained on conversational speech and then tested against transcript-dependent speech.

The Linguistic Data Consortium (LDC), in consultation with Lincoln Laboratory, NIST, and the Salk research community, created the Mixer and Transcript Reading corpora to support research, development, and evaluation of robust speaker recognition. Against this backdrop, our experimental results are given, and, finally, our conclusions and future work.

2. Mixer Read Transcription Collection

The Mixer read transcript collection [6] consisted of approximately 100 male and female speakers reading transcripts from each other’s conversations. The transcripts were created by excising 30 seconds of speech from a telephone conversation with each speaker. This resulted in over 900 transcribed utterances. Each speaker was then prompted to read the transcripts in random order over several recording sessions.
by means of an automatic system. The recordings were made over a telephone line as well as over eight auxiliary room microphones. The prompts varied in length from 1 to 49 words and included transcriptions of the disfluencies in the original speech. Some examples of prompts include:

- “Okay, now I, I, d- I know I don’t know much about the the system in America, the educa-“
- “You know, artists create music so they can share it.”
- “Wal-mart.”

The automatic system was designed so that a copy of each prompted transcript (or “prompt”) would be saved along with the approximate time marks of the speaker’s utterance. Each speaker was expected to read the entire set of transcripts correctly once. Some prompts were marked as “calibration” prompts or as “repeat” prompts. A repeat would indicate that the speaker misread the transcript and the prompting system operator would ask them to say it again. Therefore, ideally, the last “repeat” version of an utterance would be the one that was correctly read by the speaker.

It should be emphasized that the transcript-reading portion of the Mixer data set described in this paper was not thoroughly audited by the Linguistic Data Consortium (LDC). As with any data collection of this size, scope and complexity, there were many unforeseen issues. Some of the audio was not collected due to equipment failures resulting in missing data. The “repeat” labeling was not always accurate and often transcripts occurred without any accompanying audio.

Lacking the resources to audit the Mixer data ourselves, we took an automatic approach to try and overcome some of these issues. We used the 2004 BBN Byblos ASR system [3] to generate an errorful transcription of all the read data, and then performed a time-mediated string alignment to the original prompted transcripts. The word error rate of the 2004 BBN Byblos system itself on the Mixer data is somewhere between 15% and 20%. We then computed the word error rate for each instance of a given transcript and kept the version with the lowest error rate. Some transcripts were rejected completely, if the error rate was too high. The end result was a data set with 87 speakers (40 males and 30 males) where each speaker read an average of 832 out of 947 transcripts. This is a “dropout” of roughly 115 transcripts per speaker.

### 3. Experimental Setup

The significant transcript dropout mentioned Section 2 poses a problem when creating a transcript-dependent corpus. Our goal is to create an experiment with as many speakers as we can find reading the same set of utterances. Our approach to solving this problem was to create a vector representation of the data where each vector was associated with a unique prompt and each element of the vector was associated with a unique speaker. An element of a transcription vector $x$ would contain a $1$ if the speaker read the prompt and a $0$ otherwise. We then used K-means clustering with the city-block distance metric to find utterance clusters $\mathcal{S} = \{s_1, \ldots, s_K\}$ which minimize:

$$\operatorname{argmax}_S \sum_{i=1}^{K} \sum_{x \in s_i} \| x - \mu_i \|_1$$

Where $\| \cdot \|_1$ is the $L_1$ norm and $\mu_i$ is the component-wise median of the cluster $s_i$. The number of available speakers for a given cluster was then the sum of the logical “AND” of the components of the vectors in the cluster. We set a minimum threshold of 30 male and 35 female speakers for pruning the clusters. Unfortunately, getting this approach to work required setting the number of clusters to 300, which resulted in a majority of the clusters being singletons (i.e. containing only one utterance). The resulting clusters had durations ranging from 4 to 12 seconds. The number of experiments and other related statistics corresponding to the final clustering for each gender after pruning are given in Table 1. The same models were used for enrollment for all experiments, and each one was trained with a full conversation side (approximately 2.5 minutes of speech).

We also ran an experiment comparing the general effects of read speech versus conversational speech. The durations of the speech were approximately 2.5 minutes in both cases (the read speech was randomly selected). We tried using read speech in training and in test. For these experiments, we needed two sets of conversations: the first set (CONV1) was always used for train or test when read speech was used and the second set (CONV2) was always used in contrast to the read speech condition. Unfortunately, CONV2 contained speech from only 70 of the 87 speakers (40 females and 30 males).

For all our experiments, we used the Lincoln GMM/LFA system that was used in the NIST 2009 Speaker Recognition Evaluation [8]. We ran the system both with and without ZT-norm. The system had to be retrained from scratch after removing any data that came from speakers that were part of the Mixer read transcription collection (some of these speakers had been exposed in previous NIST evaluations). The $Z$-norm and $T$-norm data sets were also modified to avoid using the Mixer read transcript speakers.

### 4. Experimental Results

Figures 1 and 2 show the EER of the experiments as a function of duration. These results were obtained without using ZT-norm, which was not found to improve performance. Clearly there is a linear trend: as the durations get shorter, the EER gets worse. A plot of the residual values after fitting a linear regression to the data is also included in the bottom portion of Figures 1 and 2. As a comparison, we included a plot of the performance for conversational-style speech across the same range of durations in the figures, along with the residuals after applying a linear fit. In general, the average performance is better for conversational speech and the overall variance in performance about the mean is lower.

In order to plot a comparison of the distributions of EER for both styles of speech, we used the linear fit to remove the effects of duration and map all the data points to a nominal duration of 12 seconds. The CDFs of the EER for read and conversational speech at this nominal duration are shown in Figures 3 and 4. In general, the EERs for conversational speech are much better than for read speech (for males, 90% of the conversational ex-

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers (max)</td>
<td>198</td>
<td>198</td>
</tr>
<tr>
<td>Speakers (avg)</td>
<td>40</td>
<td>35</td>
</tr>
<tr>
<td>Target trials (avg)</td>
<td>295.2</td>
<td>295.2</td>
</tr>
<tr>
<td>Non-target trials (avg)</td>
<td>10533.4</td>
<td>10533.4</td>
</tr>
</tbody>
</table>

Table 1: Some statistics on the text dependent Mixer experiments.
<table>
<thead>
<tr>
<th>Train Style</th>
<th>Test Style</th>
<th>EER</th>
<th>DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>Conv2</td>
<td>4.60</td>
<td>2.05</td>
</tr>
<tr>
<td>Conv1</td>
<td>Read</td>
<td>4.41</td>
<td>2.13</td>
</tr>
<tr>
<td>Read</td>
<td>Conv1</td>
<td>3.50</td>
<td>1.85</td>
</tr>
<tr>
<td>Conv2</td>
<td>Conv1</td>
<td>3.59</td>
<td>1.66</td>
</tr>
</tbody>
</table>

Table 2: Performance at 2.5 minutes for read and conversational speech.

The results of these experiments suggest that there may be some degradation in performance at short durations when using text-dependent data for automatic speaker ID. This could be due to the difficulty some speakers have in reading prompted text or, in some cases, the prompts themselves may have been difficult to read (the prompts in this case were taken from other speakers’ spontaneous telephone conversations). In comparison, the performance for using spontaneous conversational speech at these durations appears to be much better and to have a lower variance across experiments.

It is interesting to note the outlier data point in Figure 1 at a duration of 7 seconds and an EER of 4.92%. This data point corresponds to an experiment with a single transcript: “but, um, I’m married now. My name is Schwartz, okay? which immediately tags me as being Jewish which I’m not, because”. Also at a duration of about 7 seconds, is the following transcript which corresponds to an EER of 12.3%: “um, I’m reading my own characteristics in the animal. Like a a pet, or, or animal used for food”. It’s difficult to say what determines the difference in speaker recognition performance for these two transcriptions, but it could be that the first one is easier to read than the second. Unfortunately we don’t currently have a way of or predicting “readability” for the transcriptions though this may be an interesting area to investigate.

A comparison of the performance of read versus conversational speech at longer durations (approximately 2.5 minutes) is given in Table 2. Swapping CONV1 and CONV2 appears to have a greater effect on performance than changing CONV2 for READ speech data. In general, using read speech for testing or training does not appear to degrade performance significantly.

5. Conclusions

Our speaker recognition results are somewhat surprising from a text-dependent versus text-independent perspective, but the conversational-transcript-dependent mode here is different and new. The results presented in Section 4 indicate that transcript-dependent test data, using conversational-speech enrollment data, can have a wide range of performances at short durations. This may be due in part to how difficult it is to for some speakers to read transcribed conversations, to how difficult some prompts are to read, or to some prompts being especially good for speaker identification. If the enrollment data were to match the test data (that is, if it was the original source speech used to extract prompts from or if it was the target speaker reading the same prompt), then one would expect the performance to be better in general but unfortunately in this case the duration of the transcripts is too short to adequately train a model. For our text-independent system, we have not seen an improvement in speaker recognition performance by choosing read English speech content versus unconstrained read speech. Furthermore, our experiments show improved speaker recogni-
tion performance by using conversational English speech over read speech processed in a text independent way. Additionally, we have seen that multiple sessions of conversational English enrollment data at a fixed target duration lead to better performance than a single session of conversational English at the same target duration. Our future work will include attempting to predict if a given transcript will result in better speaker recognition performance.

6. References


