Some Experiments on Idiolectal Differences among Speakers

1 Motivation

It is generally recognized that human listeners can distinguish between speakers who are familiar to them far better than those who are unfamiliar. This increased ability is due no doubt to speaker idiosyncrasies that are recognized by the listener, either consciously or unconsciously. These speaker characteristics offer the possibility to significantly improve automatic speaker recognition performance, if only we were able to identify and use them.

Historically in speaker recognition technology R&D, effort has been devoted to characterizing the statistics of a speaker's amplitude spectrum. And while this has included dynamic (e.g., difference spectra) as well as static information, the focus has been on spectral rather than temporal characterization. "Familiar-speaker" differences, however, surely relate to longer term speech patterns, such as the usage of certain words and phrases, and to the features tied to these patterns, such as intonation, stress and timing. The use of such patterns and features affords a promising but radical departure from mainstream speaker recognition technology.

To explore the possibility of using longer-term speech characteristics to characterize speakers, some preliminary experiments were performed using the SwitchBoard corpus. These experiments were performed in order to begin to understand and to calibrate some idiolectal differences among speakers. If such differences exist, then presumably they would exist within the context of speech patterns specific to the speakers. Therefore this study was directed toward the statistics of word sequences as a function of speaker.

2 Speaker-Dependent Language Models

N-gram language models are often used to good effect to improve speech recognition performance. These models are general models of the language, trained on very large corpora, typically including different sources from numerous speakers. And while advanced speech recognition systems usually include algorithms to adapt to different speakers, adaptation is directed largely towards acoustic (spectral) features.

It is possible to train language models for a specific speaker, of course, assuming sufficient data. The question is whether such language models are useful in distinguishing among speakers. Some preliminary experiments were conducted to explore this question using the SwitchBoard corpus.[1] These experiments were conducted to explore idiolectal differences and to comprehend the speaker characterizing potential of N-gram language models.

[1] The SwitchBoard corpus contains data from five hundred speakers collected from telephone conversations of nominally five minutes duration. The average number of conversations per speaker was eleven, and each of the conversations for a given speaker was typically on a different topic. More details may be found on the Linguistic Data Consortium's (LDC's) web site: http://www.ldc.upenn.edu/readme_files/SwitchBoard.readme.html

3 SwitchBoard Experiments

A number of experiments were conducted using the SwitchBoard corpus. All of these experiments used manual transcriptions of the speech signal as the input data. No use was made of the acoustic speech signal, per se (except as the source for the manual transcription, of course). The manual transcriptions were further processed to eliminate punctuation and transcriber comments and to add begin/end turn tags (pseudo-words). An example utterance is:

```
<start> Like uh [lipsmack] my boyfriend
listens to Guns and Roses <end>
```

Several variations of this representation might be to exclude non-lexical sounds, to ignore case, and to ignore turn boundaries. These simplifications reduce the size of the N-gram vocabulary but also reduce the richness of the representation.

3.1 Speaker Entropy

The first experiment was to compute the speaker entropy of individual N-grams. For the purpose of this study, the speaker entropy of an N-gram was defined as:

$$Entropy(Ngram) \quad = \quad -\sum_i \left\{ P_{Ngram}(Spkr_i) \cdot \log\left[P_{Ngram}(Spkr_i) \right] \right\}$$

where $P_{Ngram}(Spkr_i)$ is simply the fraction of N-gram tokens in the entire SwitchBoard corpus that were spoken by speaker i:

$$P_{Ngram}(Spkr_i) = N_{Ngram}(Spkr_i) / N_{Ngram}(total)$$

Figure 1 is a scatter plot of speaker entropy for bigrams, where entropy is plotted versus the total number of bigrams in the corpus. For infrequently occurring bigrams the low-entropy word patterns (i.e., those that are highly indicative of the speaker) include a number of speaker-specific content words. For example: "in Maryland", "South Dakota" and "Rhode Island". For more frequently occurring bigrams the (relatively) low-entropy word patterns contain a number of back-channel words. For example: "uh-huh uh-huh", "<start> Right" and "Oh <end>". There are also a number of common speech patterns that show speaker specificity and that might thus be thought of as idiolectal. For example: "in terms of", "sort of", "it were" "so forth" and "you bet". One bigram was particularly interesting in that it occurred a total of 25 times in the SwitchBoard corpus and yet had a speaker entropy of zero, meaning that it occurred only for a single speaker. This is the bigram "how shall". On further inspection this bigram was found to be part of a larger phrase, namely "how shall I say …", which occurred in half of the 26 conversations for this speaker. It is idiosyncratic speech patterns like this that we might wish to exploit in recognizing familiar speakers.

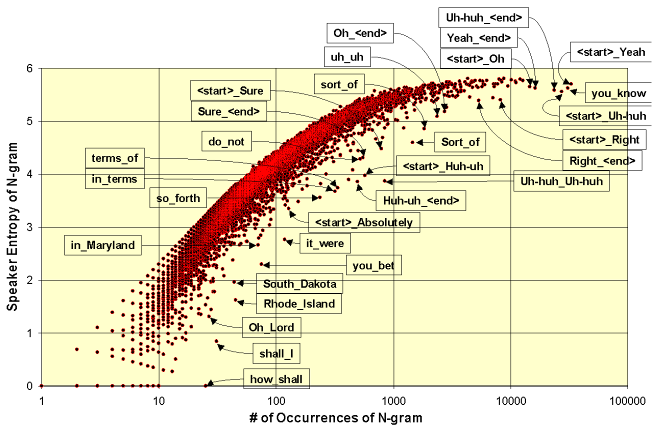

Figure 1 Speaker entropy of bigrams for the SwitchBoard corpus

3.2 Speaker Detection

Speaker detection experiments were conducted using a whole conversation side as the test segment. For each test, one true speaker conversation side was selected for the true speaker trial and one or more impostor conversation sides were selected for the impostor trials.

3.2.1 Decision Algorithms

A conventional log likelihood ratio test was used. Thus the test segment score was defined to be the log of the ratio of true speaker likelihood to background speaker likelihood for an N-gram token j, averaged over all N-gram tokens in the conversation-side:

$$Score = \frac{\sum_j \{\log[\Lambda_{TS}(j)/\Lambda_{BG}(j)]\}}{\sum_j \{1\}}$$

This formula is expressed in terms of N-gram tokens, but for efficiency the log likelihood ratio is actually computed only once for each N-gram type, k:

$$Score = \frac{\sum_k \{N_{tokens}(k) \cdot \log[\Lambda_{TS}(k)/\Lambda_{BG}(k)]\}}{\sum_k \{N_{tokens}(k)\}}$$

where $N_{tokens}(k)$ is the number of occurrences of N-gram type k in the test segment.

The N-gram likelihoods for this test were then estimated from the remaining conversations in the SwitchBoard corpus. Thus the target speaker model was created from all the conversation sides for the target speaker *except the one under test*, and the background speaker model was created from all the conversation sides in the whole SwitchBoard corpus *except those for the target speaker and the selected impostor speakers.*[1]

For most of the speaker detection experiments discussed here, target speakers were limited to those having at least 10 conversations, meaning that each target speaker model contains data from at least 9 sessions.[2] It should be noted here that in the SwitchBoard corpus each conversation was targeted to a specific topic, and that the SwitchBoard system controlled the topic selection so that no speaker (hardly) ever spoke on the same topic more than once.[3]

Figure 2 is a plot of the detection error trade-off ("DET") curve for unigrams and bigrams. Note that there is significant speaker characterizing information for both unigrams and bigrams, with bigrams providing the best performance.

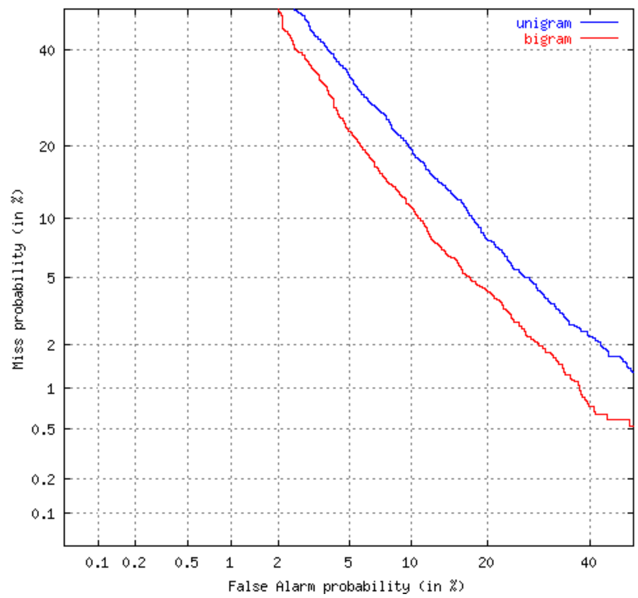

Figure 2 Speaker detection performance on conversation sides for unigram and bigram likelihood ratio scores

Considering the statistical correlation between recurrences of the same N-gram, the score formula was modified to discount multiple occurrences of the same N-gram in a test segment:

$$Score = \frac{\sum_k \{N_{tokens}^{1-D}(k) \cdot \log[\Lambda_{TS}(k)/\Lambda_{BG}(k)]\}}{\sum_k \{N_{tokens}^{1-D}(k)\}}$$

where D is the discount factor, with permissible values of D between 0 and 1.

For $D = 0$ there is no discounting of N-gram tokens, and for $D = 1$ there is complete discounting. (With complete discounting, a particular N-gram type will contribute the same increment to the score *regardless of how many times that N-gram occurs during the test segment.*)

Figure 3 is a DET plot that compares speaker detection performance with and without discounting of N-gram tokens. Note

[1] These estimated N-gram likelihoods were smoothed by adding 0.001 to each likelihood estimate.

[2] This reduced the total number of target speakers to 217.

[3] The fact that speakers discussed a different topic during each conversation is significant because this implies that the speaker detection performance is *not* attributable to the topic. On the contrary, the speaker detection performance is *despite* changes in the topic. (Only four percent of the speakers ever spoke on the same topic more than once, and well under one half percent of all conversations were on a repeated topic for a speaker.)

that discounting degrades performance for unigrams but improves performance for bigrams. Figure 4 shows DET performance for several values of discounting for bigrams. Note that the best performance for bigrams is obtained with complete discounting. Therefore, complete discounting will be used for the remainder of the experiments discussed in this paper.

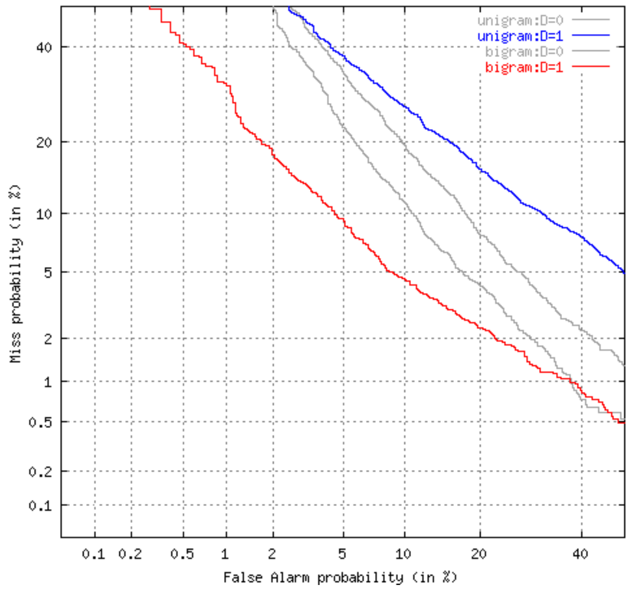

Figure 3 Speaker detection performance for unigrams and bigrams with and without token discounting

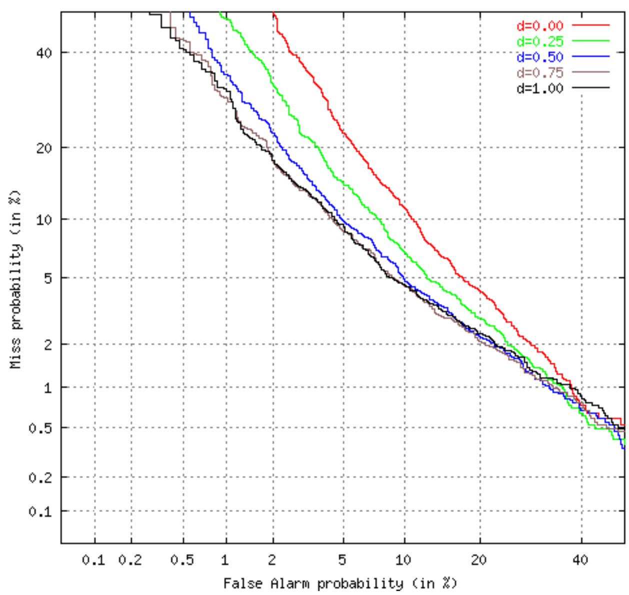

Figure 4 Speaker detection performance for bigrams as a function of token discounting

3.2.2 Reduced N-gram Representations

The N-grams, taken from the original SwitchBoard transcriptions, preserve information beyond that provided by basic SNOR-style transcriptions. Specifically, SwitchBoard transcripts also provide upper/lower case information, non-word sounds (specifically sounds described by bracket-enclosed keys such as “[laughter]” and “[lipsmack]”), and turn start/end tags. Figure 5 shows the effect on speaker recognition performance of eliminating these components.

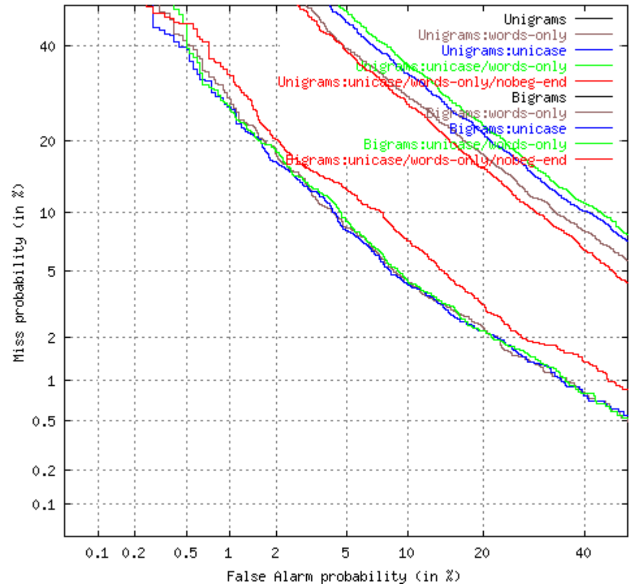

Figure 5 Speaker detection performance for N-grams with reduced representation

For both bigrams and unigrams, there is no change in performance from excluding non-words. Beyond this, however, the effect is quite different for unigrams and bigrams. For unigrams, eliminating case information degrades performance significantly. And additional elimination of non-word information degrades performance further (curiously, since there is no degradation in performance when non-words alone are eliminated). For bigrams, there seems to be no effect on performance, regardless of the presence or absence of case and non-word information. This might be due to an ability of bigrams to (redundantly) represent the information conveyed by case. Finally, the effect of eliminating turn information is opposite for unigrams and bigrams. For unigrams, the elimination of turn information (represented by single “<start>”/ “<end>” tokens) gives the best performance, while for bigrams elimination of turn information gives the worst performance.

3.2.3 Performance versus Amount of Test Data

It would be interesting to understand how performance varies with the amount of test data. To assess this aspect of performance, a scatterplot of bigram test scores is shown in figure 6, where each test score is plotted versus the number of bigram tokens in the test segment. Overlaid on this scatterplot are plots of the mean values and standard deviations of test scores for subsets of scores divided according to number of bigram tokens. Perhaps more relevant is the derivative F-ratio measure, which shows a sharp rise with the size of the test segment. Note also that there is no suggestion that the F-ratio might be approaching an asymptote, up to the limits imposed by the SwitchBoard corpus.

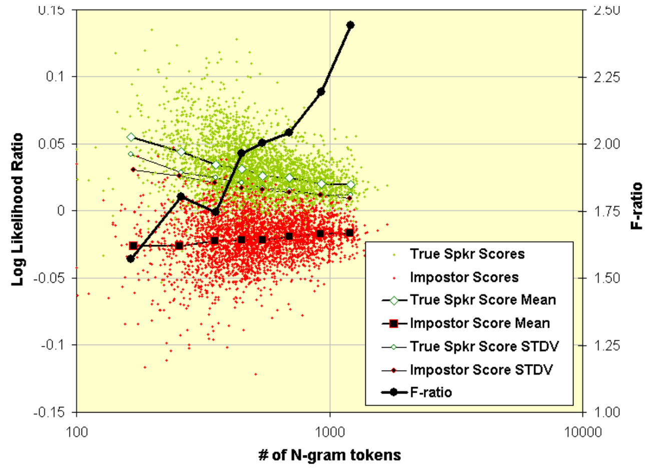

Figure 6 Scatterplot of speaker detection scores for bigrams as a function of the number of bigram tokens in the test segment.

3.2.4 Performance versus Number of Training Sessions

It seems surprising that a speaker-dependent N-gram language model, trained on a rather small number of short conversations, could provide the level of speaker detection performance that has been observed. Certainly this supports the notion of idiolect – speaker-specific usage of words and phrases. Nevertheless, it would seem that a significant amount of training data would be required to adequately calibrate idiolect for speaker recognition.

To gain some understanding of how performance varies as a function of the amount of training data, the target models were partitioned into different subsets according to how many conversation sides were used in creating the target model. Results are shown in Figure 7 for bigrams. While there exists a modest level of speaker detection performance for even a single training session, performance climbs steadily up to the limit imposed by the SwitchBoard corpus, with each doubling of training data resulting in approximately a halving of error rate.

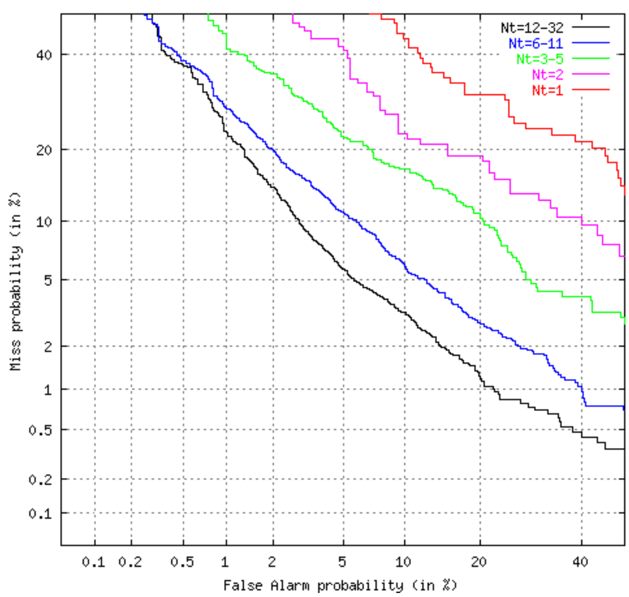

Figure 7 Speaker detection performance as a function of the number of training sessions for bigram models.

3.2.5 Performance of Low- and High-frequency Bigrams

To gain some understanding of the source of the speaker characterizing power, two experiments were run to progressively prune away first the low-frequency bigrams and second the high-frequency bigrams. This pruning was according to the total number of bigrams occurrences for the entire SwitchBoard corpus. Figure 8 is a DET plot showing the effect of excluding the low-count bigrams, and figure 9 is a DET plot showing the effect of excluding the high-count bigrams.

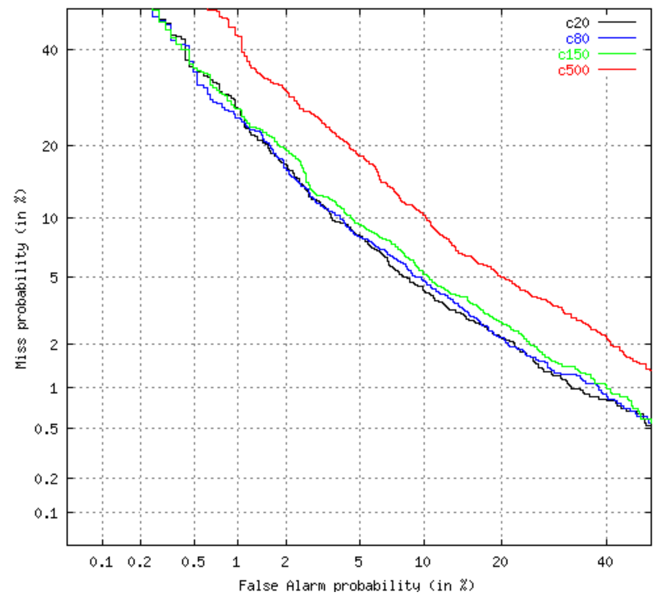

Figure 8 Speaker detection performance excluding low frequency bigrams

Note that there is little effect of excluding low-count bigrams up to a count of 150. This is encouraging, because there are only 2500 bigram types with a count 150 or more, which account for half of all bigram tokens. (A cumulative distribution of unigram and bigram types and tokens for the SwitchBoard corpus, versus frequency of occurrence, is given in figure 10.)

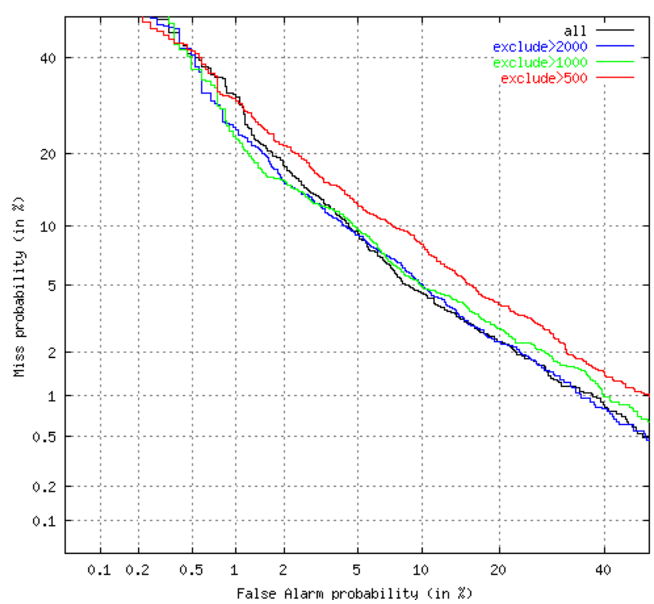

Figure 9 Speaker detection performance excluding high frequency bigrams

For high-count bigrams, there is little effect down to a count of 1000. This accounts for fewer than 300 bigram types, but over one quarter of all bigram tokens. So, it appears that most of the idiolectal action, at least with respect to the use of bigrams for speaker recognition, is in the third most likely quartile of bigrams.

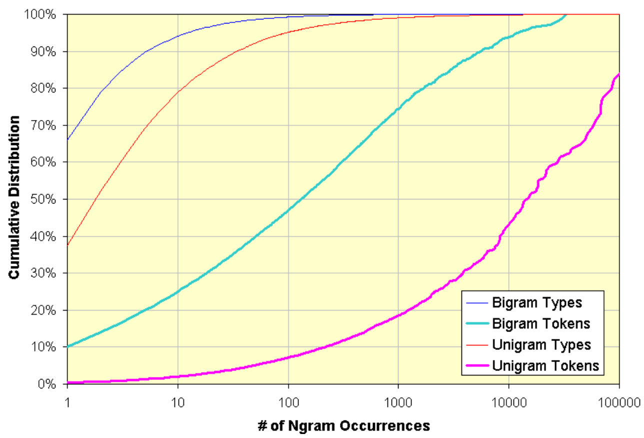

Figure 10 Cumulative distribution of N-gram types and tokens

3.2.6 Demographic Factors that Affect Performance

There is a clear distinction in the acoustics between male and female speakers, which is not present in the transcription of course. There may, nonetheless, be consistent idiolectal differences between men and women that are exhibited in the speaker detection task. This is affirmed in the contrast between same-sex and cross-sex speaker detection performance shown in the DET plots in figure 11. Curiously, there seems to be little or no difference between same-sex and cross-sex performance for female models, while the difference is a striking factor of 4 for male models.

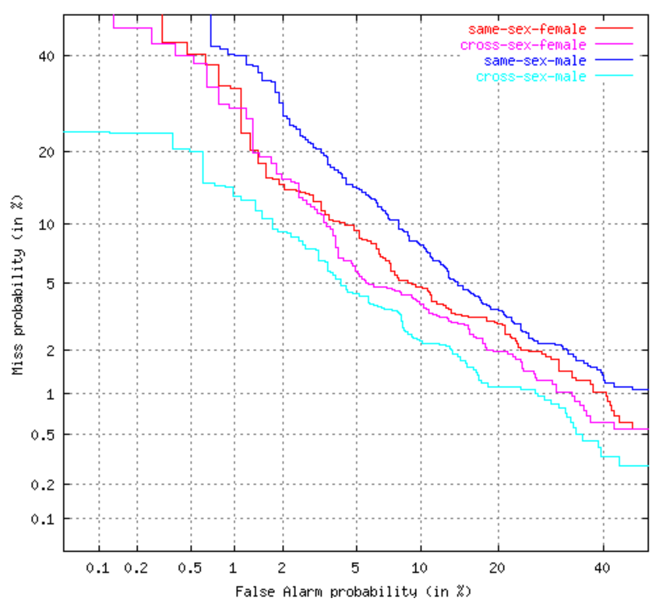

Figure 11 Comparison of speaker detection performance for same-sex versus cross-sex impostors

Another factor of perhaps only academic interest is the significance of age difference between impostor and target. To assess this, a scatterplot of impostor score versus age difference is presented in figure 12. While there is no apparent trend visibly obvious in the scatterplot itself, a second order polynomial regression line shows that impostor scores do tend to become worse as the age difference between impostor and target increases. Several speculative explanations for this phenomenon are possible. For example, there may be stage-of-life factors that influence a speaker's idiolect. Or this may be a side effect of the evolution of language. Or this effect may be a mere statistical anomaly.

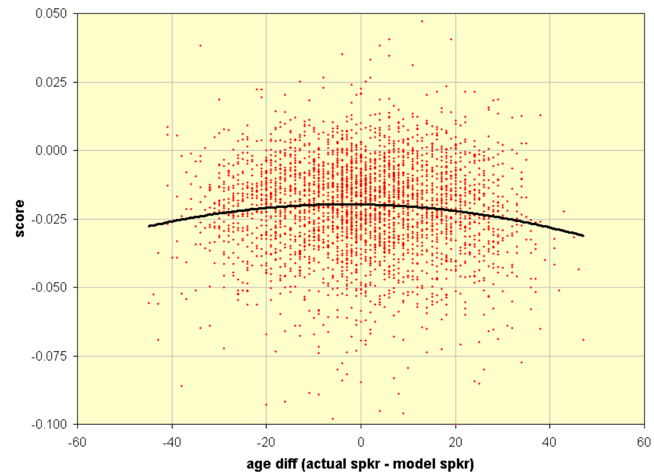

Figure 12 Scatterplot of impostor scores versus age difference

4 Conclusions and Recommendations

The performance of speaker detection based upon bigram statistics is surprisingly good, at least for the SwitchBoard corpus as studied. Surprising from several aspects, not just that speaker detection error rates are low:

- Although performance was observed to continue to improve as the amount of training data was increased, nonetheless good performance was observed for a surprisingly small number of training conversations.
- Performance was maintained while excluding all but a small number of bigrams, on the order of a few thousand. These bigrams are namely those that occur most frequently. (This helps to explain why it is that good performance is achieved with a relatively small amount of training data.)

These experiments are very encouraging. They suggest that it may be feasible to exploit "familiar speaker" characteristics with a reasonable amount of training. They also suggest that it might be reasonable to create a technology that (automatically) finds the needed higher-level speech patterns (because they occur with sufficient frequency to exhibit multiple occurrences in the training data).

Further exploration of these ideas seems likely to produce technology of great value for speaker recognition applications and certainly of great scientific merit. One of the most promising areas would seem to be in exploiting the synergy between a speaker's language and acoustic characteristics. This can be done by more than simply combining language and acoustic scores. Rather, it may well be far more discriminative to condition the acoustic calibration of a speaker on those speech patterns specific to that speaker's idiolect.