# Speaker Recognition based on Idiolectal Differences between Speakers

*George Doddington*

National Institute of Standards and Technology, USA
doddington@nist.gov

## Abstract

"Familiar" speaker information is explored using non-acoustic features in NIST's new "extended data" speaker detection task.[1] Word unigrams and bigrams, used in a traditional target/background likelihood ratio framework, are shown to give surprisingly good performance. Performance continues to improve with additional training and/or test data. Bigram performance is also found to be a function of target/model sex and age difference. These initial experiments strongly suggest that further exploration of "familiar" speaker characteristics will likely be an extremely interesting and valuable research direction for recognition of speakers in conversational speech.

## 1. Introduction

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.

Similar work has been performed on the recognition of authors. Probably the best know of these is the work of Mosteller and Wallace on determining authorship of the Federalist papers.[2] There is reason to hope that speech, being less constrained by convention, might support even greater speaker characterizing power.

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 information (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, NIST has added a new "extended data" speaker detection task to its yearly evaluation plan. This paper reports on some preliminary experiments that were conducted within this new task. 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 exists. The question is whether such language models are useful in distinguishing among speakers. To answer this question, some preliminary experiments were conducted using the SwitchBoard corpus.[3] These experiments were conducted to explore idiolectal differences and to comprehend the speaker characterizing potential of N-gram language models.

## 3. SwitchBoard Experiments

A number of experiments were conducted using the SwitchBoard corpus. In all of these experiments, the input data were manual transcriptions produced by ISIP.[4] 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 [noise] 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 number of N-grams to be dealt with, but they 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:

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

where $P_{Ngram}(Spkr_i)$ is 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 information versus frequency of occurrence for bigrams, assuming a uniform prior distribution of speakers. (There are 520 speakers in the SwitchBoard corpus.) Note the examples of the relatively informative bigrams. These speaker-informative bigrams are largely unrelated to content and might be described as stylistic. For example: "you bet", "it were", "so forth", "in terms of", "uh-huh uh-huh", "<start> sure" and "yeah <end>". 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.

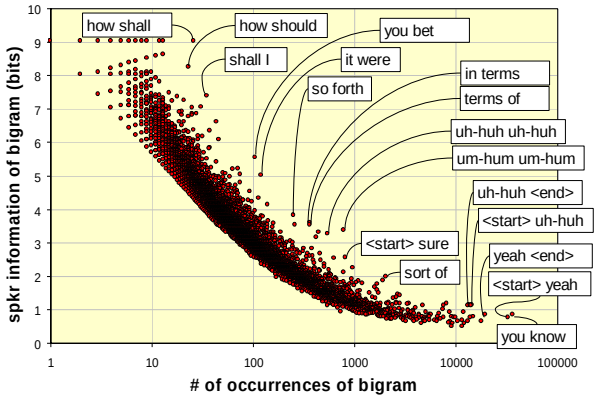


*Figure 1* Speaker information contained in word bigrams, tabulated over the whole SwitchBoard corpus

### 3.2. Speaker Detection

Speaker detection experiments were conducted according to NIST's extended data speaker detection task.[1] Each test used a whole conversation side as the test segment, with the model being trained on from 1 to 16 conversation sides.

#### *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. In order to smooth the log likelihoods, a value of 0.001 was added to each likelihood before taking the logarithm.

The N-gram likelihoods for this test were estimated from the conversation sides specified in the control file for the extended data task. Thus the target speaker model was created from a selected training set of conversation sides for the target speaker. The number of conversation sides used to train the target model was constrained to be either 1, 2, 4, 8 or 16. The background model was computed from a set of over 400 speakers drawn from the SwitchBoard corpus who were not used for testing. Six different background models were defined and, using a jackknife procedure, all conversation sides in the SwitchBoard corpus were used for test.

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. This is very helpful in avoiding a content bias that might boost speaker recognition performance artificially based on topic selection. (There was evidence of a residual bias of this kind, however, based on the discovery of speaker-informative bigrams such as "in Maryland" and "Rhode Island".)

Figure 2 is a detection error trade-off ("DET") plot for unigrams and bigrams, using 8 training conversation sides in the target model. Note that there is significant speaker characterizing information for both unigrams and bigrams, with bigrams providing the best performance.

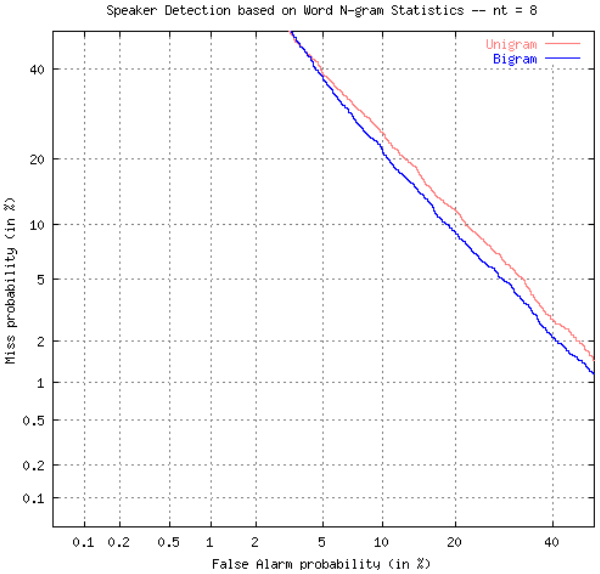


*Figure 2* Speaker detection performance for unigram and bigram likelihood ratio scores

### 3.2.2. *Performance as a function of bigram frequency*

To gain some understanding of the source of the speaker characterizing power, an experiment was performed to progressively prune away the low-frequency bigrams. This pruning was according to the total number of bigram occurrences for the entire SwitchBoard corpus. Figure 3 is a DET plot showing the effect of excluding the low-count bigrams. It is notable that performance actually improves as the low-frequency bigrams are eliminated. A pruning threshold of 200 gives the best performance. At this pruning threshold, all but the most frequently occurring 2000 bigrams are eliminated. This is further evidence that it is a speaker's commonly used word patterns that have discriminative power.

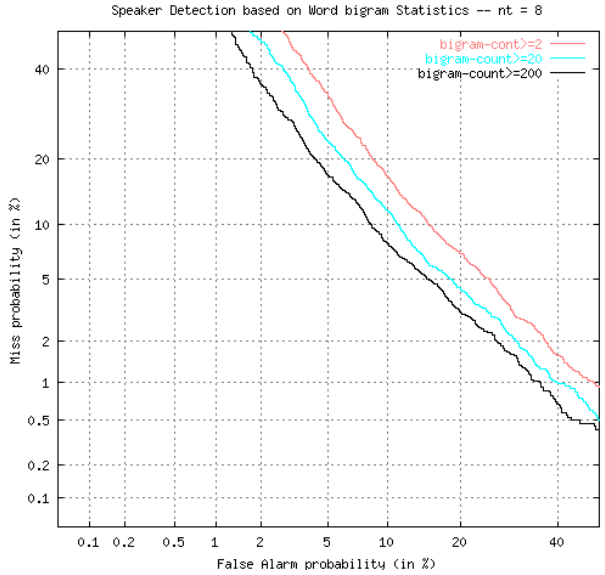


*Figure 3* Speaker detection performance for bigrams, excluding those bigrams that occur infrequently, namely bigrams with counts of less than 2, 20 and 200 over the background model training data

### 3.2.3. *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 4 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.

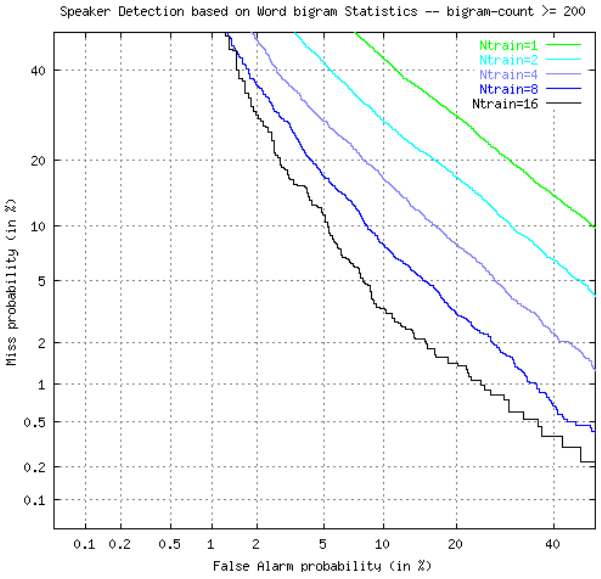


*Figure 4* Speaker detection performance for bigrams, as a function of the number of conversations used to train the target model, namely for 1, 2, 4, 8 and 16 training conversations

### 3.2.4. *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 5, 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.

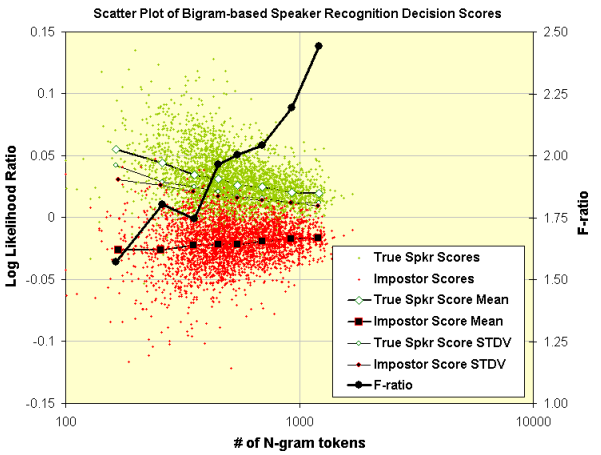


*Figure 5* Scatterplot of speaker detection scores for bigrams as a function of the number of bigram tokens in the test segment

### 3.2.5. *Demographic Factors that Affect Performance*

There is a clear distinction in the acoustics between male and female speakers. A natural consequence of this is that speaker recognition systems perform far better in discriminating between opposite sexes than same sexes. This acoustical contrast 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 6.

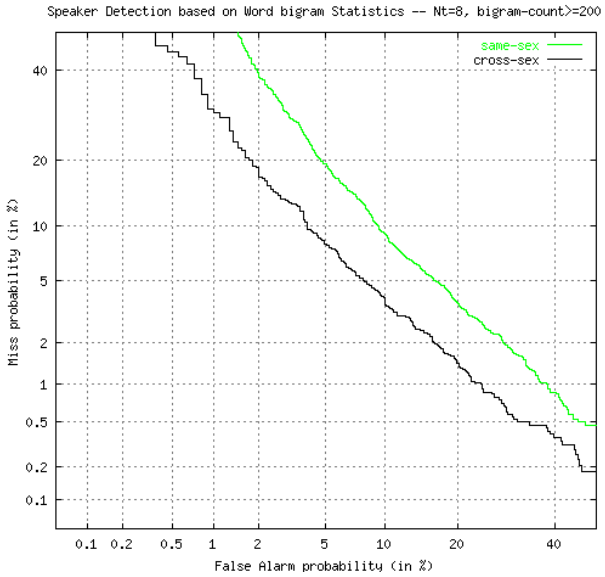


*Figure 6* Speaker detection performance for bigrams for same-sex impostors versus opposite-sex impostors

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 7. 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.

## 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.)

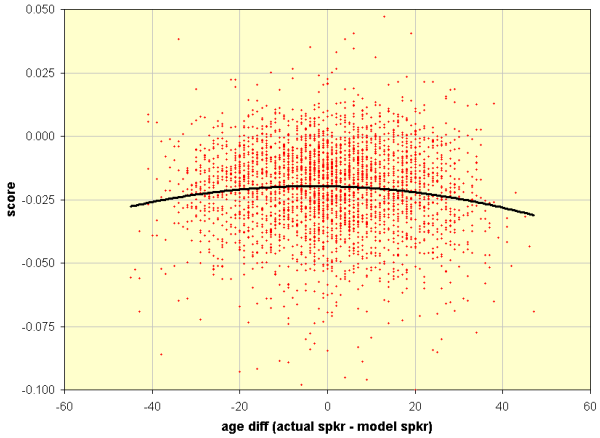


*Figure 7* Scatterplot of impostor scores versus age difference between the impostor and the model speaker

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.

## 5. References

[1] The extended data task definition is included in NIST's 2001 evaluation at www.nist.gov/speech/tests/spk/2001/

[2] *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Frederick Mosteller and David Wallace, 2nd Edition of *Inference and Disputed Authorship: The Federalist.* Springer-Verlag, New York, 1984

[3] The SwitchBoard corpus is described in more detail at www.ldc.upenn.edu/readme_files/switchboard.readme.html

[4] The ISIP transcriptions and associated documentation are available at www.isip.msstate.edu/projects/switchboard