# GENDER-BASED LANGUAGE MODELING: SALIENT FEATURES AND INTERPOLATION METHODS

A Thesis

Presented in Partial Fulfillment of the Requirements for

the Distinction Designation in the College of Engineering Bachelor of Science

of the Ohio State University

By

Annatala Trixie Wolf, B.A. Psychology

* * * * *

The Ohio State University

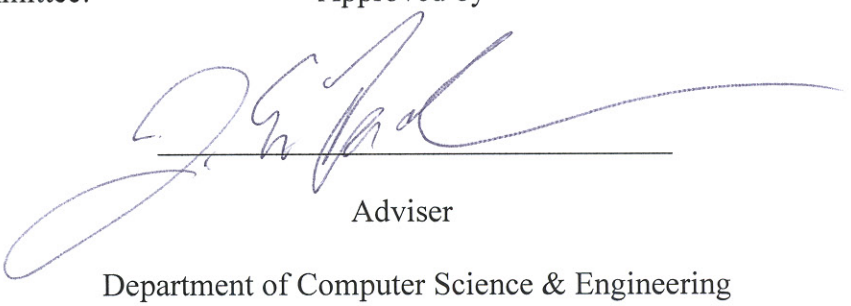
2006

Engineering Honors Thesis Committee:

Dr. Eric Fosler-Lussier, Adviser

Dr. Donna Byron

Approved by



Adviser

Department of Computer Science & Engineering

ABSTRACT

Sociolinguistic studies suggest that a relationship exists between the gender of a speaker and the words she or he chooses in conversational speech and writing (1, 7, 8). However, little research exists within computational linguistics on the subject of using the classification of gender as a tool for predicting word choice in automatic speech recognition systems. Our study presents an analysis of a recent study by Boulis and Ostendorf (2005), which demonstrated success in using words as an automatic classification of speaker gender, but failed to significantly improve perplexity in automatic speech recognition by using gender-based models (1). Specifically, small but insignificant gains over a general, non-gendered training model appeared when gender-based language models were interpolated with the general model. In our study, we replicate parts of the Boulis and Ostendorf study. We then use a chi-square test to define a different measure of topicality than the one used by Boulis and Ostendorf, and then extend this measure to produce a nonlinear interpolation of the same language models. While the results of each of our experiments are consistent with the conclusions of the Boulis and Ostendorf study, avenues for exploring possible further interpolation techniques remain open.

# INTRODUCTION

Recent research in sociology and linguistics has shown relationships between gender and word choice in formal and conversational writing (7, 8), as well as in conversational speech, such as telephone conversations (1). For one example, the word "dude" in conversational speech is far more prevalent among male speakers (9). Despite these relationships, there exists little available research in computational linguistics on the subject of gender as a tool for the statistical modeling of word choice.

The recent (2005) study by Boulis and Ostendorf, "A Quantitative Analysis of Lexical Differences Between Genders in Telephone Conversations" (1), uses the Fisher corpus of phone conversations (2) to examine this topic. Specifically, the Boulis and Ostendorf study empirically demonstrates the ability to use gendered differences in conversational speech as a basis for gender-based document classification, with some features in speech being more discriminative for this purpose than others. In their study, classification of male-only versus female-only conversations was almost perfectly accurate, while classification of mixed-gender conversations was less accurate. Classifying the gender of one speaker based on the transcript of the other speaker was also above chance level, but only by a small margin. Boulis and Ostendorf further identified the major discriminative lexical features associated with both genders, and demonstrated that the majority of these features were content-bearing words (1). Many of these discriminative words (such as swear words and proper names), however, were

uncommon in the corpus, and less likely to be useful for automatic speech recognition models. While uncommon features may be useful for adding statistical weight to a model predicting class (gender) based on speech, they are less useful for modeling the converse relationship of predicting speech based on class. This is because uncommon features do not appear frequently enough in speech for an increase in their predictability to greatly affect the accuracy of speech prediction.

Boulis and Ostendorf briefly asked whether the converse is true: if words can predict gender, can gender help automatic speech recognition systems predict word choice, thus reducing perplexity in automatic speech recognition systems? A brief exploration of gender-based language modeling revealed that gendered models were less effective than a general combined model, presumably because more training data was available with the combined model. Boulis and Ostendorf found that a linear interpolation of each gendered model with the combined model did produce gains over the combined model, but these gains were not significant (1).

Our study replicates in part the design of the Boulis and Ostendorf study, confirming their results on the same corpus. In an attempt to extend the study and look for possible opportunities to exploit gendered lexical differences in automatic speech recognition systems, we compare a different method for evaluating the discriminative power of features and use this measure as a basis for performing our own set of interpolations. While our interpolations fail to produce a measurable improvement in the performance of automatic speech recognition systems, several avenues remain open for future tests of non-linear interpolation models.

## METHODOLOGY

The Fisher corpus represents a large body of conversational speech tagged with the gender of both speakers (3). It consists of a series of transcribed telephone conversations between randomly assigned speakers, with each speaker engaging in five or fewer separate calls. For each phone call, the speakers were assigned a topic. The topicality of each call is rated from 0 to 4 by a transcriptionist, with 4 being most on-topic. Boulis and Ostendorf used a version of the Fisher corpus containing just over 12,000 telephone conversations, with conversations removed if one of the speakers was non-native or if the topicality was marked as level 0 or 1 (off-topic). Punctuation was removed (usually absent, it is applied somewhat inconsistently across the corpus), acronyms were combined, and each of the transcription files was separated into two conversation sides. Boulis and Ostendorf also converted word fragments to a single "wordfragment" token, presumably to consolidate these features in the case that false-starts and interruptions might be generally correlated with gender. "Topic bias" was removed by matching the same proportion of conversations by gender, and eliminating (at random) the extra conversations. This was done to limit the likelihood that document classification might be classifying topics rather than gender, in the event that one or more topics in the random distribution occurred more frequently among one gender or the other. "Speaker bias" was removed by ensuring that the data was separated such that each speaker appeared only in the training set or only in the testing set. Rainbow, a

statistical toolkit for creating document classification models, was used as a document classifier (2). A document classifier is a tool that can be used to assign an electronic document into one or more categories, a common problem in information science. Rainbow has the ability to both build and test models for document classification. Here, the "classes" were the speaker and listener gender. The SRILM toolkit was used for the language model tests (4).

The experiments in our study occurred in two stages. In the first stage, the document classification experiments were replicated on a subset of the corpus used by the authors. We originally had access to a restricted subset of the Fisher corpus with less than half as many documents as the version used by Boulis and Ostendorf. The smaller set meant less training data overall and a larger baseline perplexity in our measurements. The elimination of topic bias, which depends on document matching in 40 categories, required a larger percentage of documents be discarded from our study than from the original, which compounded the size difference further: our study used a total of 4997 conversation sides for training and 1249 for testing, while the original study used 14969 for training and 3738 for testing. The overall difference in the size of the training data is probably the largest factor contributing to the general size discrepancy between our results and the results obtained by Boulis and Ostendorf, but the relative values between various models are still comparable. Our initial replication omitted two of the preparations made by Boulis and Ostendorf, since replication was already inexact from the size of our training data. Specifically, word fragments were not consolidated and speaker bias was not removed for the initial tests on document classification.

In the second stage, separate features were drawn from the training data for comparison with the features reported by the Boulis and Ostendorf study, and the language model tests were performed. During the second stage, we had access to the full corpus used by Boulis and Ostendorf, but once non-English speakers and topic bias were considered, we were left with 13852 conversation sides for training and 3374 for testing: 1117 conversations fewer than the data set reported by Boulis and Ostendorf. This discrepancy was unexpected, because Boulis and Ostendorf's description of the methodology used to match topicality within genders was quite clear.

Due to our interest in exploring the results of various nonlinear interpolation methods for class-dependent bigram models, one-third of the testing section was removed into a special category for development data (1124 sides, 607 female and 517 male). For the chi-square analysis and all of the speech recognition tests, word fragments were consolidated and speaker bias was removed.

Our replication tested several of the experiments performed by Boulis and Ostendorf, using similar methods on our smaller subset of the Fisher corpus. Most of these experiments entailed using the Rainbow document classifier to predict gender from word choice. The details of the individual experiments are left to the results section. On the topic of discriminative features, we compared Boulis and Ostendorf's method using Kullback-Liebler distance with three of our own: information gain (also referenced by Boulis and Ostendorf), chi-square effect size for document frequency, and chi-square effect size for feature frequency.

The chi-square test is a statistical method for determining the likelihood that a difference in occurrences between two groups is due to chance. It may also act as a

measure of effect size for topicality, as or in our case, in-gender, as some authors have proposed for use with nonlinear interpolation models (6). The formula for chi-square values is calculated as follows for each feature in the corpus:

**(observed frequency – expected frequency)$^2$ / expected frequency**

These values are calculated for each of four categories, and all four categories summed to get a final, single value for the feature. The categories are: feature frequency in male gender, "other" frequency in male gender, feature frequency in female gender, and "other" frequency in female gender. Here, "other" frequency refers to the total number of features minus the frequency of the feature in question. Both unigram and bigram chi-square values can be calculated separately. Additionally, the observations can be made on a term frequency basis (how many times a unigram or bigram appears versus the appearance of other unigrams or bigrams), or on a document frequency basis (how many documents contain a given word versus how many do not).

In both cases, unigram and bigram chi-squares were calculated against only features of the same type. This ensured that the probability of each set of features would only be weighed against its own kind, since unigrams and bigrams do not occur with equal frequency. The different chi-square values were then used as a measure of the effect size for significance of a word being on-topic.

The experiments on the effect of feature selection criteria on gender classification by vocabulary size and exploration of content-bearing words were not replicated, but similar topics are discussed in the results section.

A number of experiments were run to determine if gender lexical differences could be used to enhance automatic speech recognition, most of these dealing with

nonlinear manipulation of word counts used to produce bigram language models, or modifications to the probabilistic language models themselves. These experiments are discussed in greater detail in the results section.

## RESULTS

The first stage of our results concerns replication of the Boulis and Ostendorf classification experiments on a smaller set of data. The first experiment in the Boulis and Ostendorf study tested various learning methods of the Rainbow document classifier to see if gender classification was possible, and compared the classification accuracy of various learning methods with the Rainbow toolkit: Cosine, Naïve Bayes, Maximum Entropy, and Support Vector Machine (SVM). This demonstrated which method was most effective in creating a model from conversational training data that could predict gender based on the features in a separate transcribed conversation. Features tested included individual words (unigrams), and combination of unigrams and word pairs (bigrams). Unigram and bigram document classification models were built on male and female data for each of the learning methods. The results are illustrated in Figure 1.

Wolf study

| Method | Unigram | Bigram |
|---|---|---|
| Cosine | 74.0% | 79.9% |
| Naïve Bayes | 78.0% | 79.5% |
| MaxEnt | 82.0% | 80.9% |
| SVM | 81.2% | 82.9% |

Boulis and Ostendorf study

| Method | Unigram | Bigram |
|---|---|---|
| Cosine | 76.3% | 86.5% |
| Naïve Bayes | 83.0% | 89.2% |
| MaxEnt | 85.6% | 90.3% |
| SVM | 88.6% | 92.5% |

Figure 1. Classification accuracy of Rainbow learning methods.

Our results were comparable to the original study, with SVM being the most successful language model overall. The lower accuracy of our results across the board is most likely an effect of our limited amount of training data.

The second experiment produced a confusion matrix for classification of gender, based on the results from a single 4-way model using unigram features and an SVM learning method. For each category, the initial letter indicates the gender of the speaker, and the subsequent letter indicates the gender of the listener. The rows represent the actual category, and the columns are the classifier's prediction of gender. This is shown in Figure 2.

Wolf study

| | FF | FM | MF | MM | F-measure |
|---|---|---|---|---|---|
| FF | 212 | 74 | 27 | 32 | .594 |
| FM | 106 | 74 | 66 | 34 | .290 |
| MF | 32 | 53 | 95 | 100 | .360 |
| MM | 19 | 28 | 60 | 237 | .635 |

Boulis and Ostendorf study

| | FF | FM | MF | MM | F-measure |
|---|---|---|---|---|---|
| FF | 1447 | 30 | 40 | 65 | .778 |
| FM | 456 | 27 | 43 | 77 | .074 |
| MF | 167 | 25 | 104 | 281 | .214 |
| MM | 67 | 44 | 210 | 655 | .638 |

Figure 2. Confusion matrices for 4-way classification (unigrams, SVM).

F-measure is the weighted harmonic mean of precision and recall, and so it measures the performance of how well retrieved information matches the actual information:

**F = 2 * precision * recall / (precision + recall)**

Precision is the percentage of documents retrieved which are relevant: here, the proportion of correctly gender-matched documents to the proportion of documents mislabeled to that gender. Recall is the percentage of relevant documents that are retrieved: here, the proportion of correctly gender-matched documents to the proportion of documents that were not properly matched to that gender.

The results are comparable across both studies, with F-measure being highest for same-gender conversations. FM and MF are confused with FF and MM, respectively, which can be mapped to single gender detection. The smaller sample set likely accounts for the majority of the variation in our statistics.

This information led to the third experiment, which compared classification accuracy in same-gender versus cross-gender samples. The results compare four new classification models selected to represent the samples, and are summarized in Figure 3. Here, FF-MM is the classification of female-only versus male-only conversations, while FM-MF is the classification of mixed-gender conversations by the speaker's gender.

| Wolf study | | | Boulis and Ostendorf study | | |
|---|---|---|---|---|---|
| Categories | Unigrams | Bigrams | Categories | Unigrams | Bigrams |
| FF-MM | 87.57% | 88.44% | FF-MM | 98.91% | 99.49% |
| FM-MF | 68.86% | 74.38% | FM-MF | 69.15% | 78.90% |

Figure 3. Classification accuracies in same-gender and cross-gender conversation.

The results are comparable across both studies, clearly illustrating the increased confusion in classifying conversations of mixed gender on the basis of the speaker's word selection. As with previous experiments, the reduction in accuracy in our study is likely due to the smaller amount of available training data.

The next experiment examined to see if the gender of one speaker could be determined only from the other speaker's transcript. This is summarized in Figure 4.

| Wolf study | | | Boulis and Ostendorf study | | |
|---|---|---|---|---|---|
| | Unigrams | Bigrams | | Unigrams | Bigrams |
| FF-FM | 59.65% | 61.40% | FF-FM | 57.94% | 59.66% |
| MM-MF | 66.51% | 64.27% | MM-MF | 60.38% | 59.80% |

Figure 4. Classification accuracies by gender of opposing speaker.

Our results were again comparable to the original study. The classification was slightly above chance margin in all cases, though our results were surprisingly better. While this may be due to variations in the sample from our restricted training and test sets

during this first stage of replication, the possible effect of the inclusion of individual word fragments was also a potential source of this deviation.

The fifth experiment examined gender-based features, to see if the classification accuracy came largely from a small subset of features, or from the larger vocabulary. Boulis and Ostendorf measured smaller and smaller subsets of the most discriminative features in a vocabulary, in an attempt to determine how many features are responsible for classification accuracy. Both Kullback-Liebler smoothed distance and information gain were used to form two separate classes of models with restricted vocabularies for classification (70%, 40%, 10%, and 3% of the total features, apparently arbitrary choices). Unigram and bigram features were used with each of the models. Due to technical limitations in getting the Kullback-Liebler smoothing in Rainbow to run properly on our system in limited time, we replicated only the information gain models, and then only on the 70%, 40%, and 10% restrictions. We compare these models in Figure 5.

Wolf study

| | 1.0 | 0.7 | 0.4 | 0.1 |
|---|---|---|---|---|
| IG-1 | 81.2% | 82.4% | 82.3% | 79.8% |

Boulis and Ostendorf study

| | 1.0 | 0.7 | 0.4 | 0.1 |
|---|---|---|---|---|
| IG-1 | 88.6% | 88.5% | 88.9% | 87.6% |

Figure 5. IG-pruned trend for classification accuracy for unigram features.

The general trend is similar across both studies. The best classification in both comes from using around 40% of the features, while a slight reduction in classification accuracy appears at around 10% of features—but this reduction is not highly significant.

For KL-distance, Boulis and Ostendorf reported that the best results occurred at around 70% of features, because for the KL measure a performance hit first appeared at 40%. They also found that the number of "irrelevant" features was small, about 5000 features. This was determined by using smaller and smaller subsets of the least-relevant unigrams until classification accuracy dropped to near-chance levels (below 60%).

Boulis and Ostendorf also reported the most discriminative features for both genders, using Kullback-Liebler distance. Our comparison used information gain instead, for the same technical reasons.

| Wolf study | | | Boulis and Ostendorf study | | |
|---|---|---|---|---|---|
| Rank | Male | Female | Rank | Male | Female |
| 1 | wife | husband's | 1 | dude | husband |
| 2 | shit | gosh | 2 | shit | husband's |
| 3 | hundred | goodness | 3 | fucking | refunding |
| 4 | basically | daughter | 4 | wife | goodness |
| 5 | gotta | pretty | 5 | wife's | boyfriend |
| 6 | wife's | mhm | 6 | matt | coupons |
| 7 | guy | children | 7 | steve | crafts |
| 8 | how're | boyfriend | 8 | bass | linda |
| 9 | david | mom | 9 | ben | gosh |
| 10 | guys | baby | 10 | fuck | cute |

Figure 6. Top discriminative features for classification of gender.

The general theme of the words is consistent across both studies, with some of the same words present in both lists. The Boulis and Ostendorf list included a few proper names, and the discovery of "david" not included in the Boulis and Ostendorf list reinforced the idea that names themselves are discriminative features. This was possibly because people introduce themselves during the conversation, although a brief inspection of the corpus also showed a number of instances where names were discussed within the context of pop-culture: male-correlated "Ben" referred to "Ben Affleck", "Ben Stiller", and "Ben & Jerry's ice cream". Male-typed and female-typed exclamations such as invectives and "goodness" appear in both lists, and relatives and family words were again associated with female speech. Reference words for other-gender relationships such as "boyfriend" and "wife" were naturally correlated with the opposite gender in both studies.

The second stage of experiments began with the intent to see if manipulation of gender-based language models could reduce perplexity in automatic speech recognition systems. For this purpose, it was important to identify which features are discriminative in the other direction: whereas the Boulis and Ostendorf study identified features that were discriminative for the classification of documents by gender, we were interested in determining which features were directly discernible from gender. To do this, a chi-square analysis was performed on features both for term frequency and document frequency. For term frequency, the chi-square value represented the statistical significance of the variation in appearances of term by speaker gender. For document frequency, the chi-square value represented the statistical significance of the variation in number of documents containing the term by speaker gender.

This analysis methodology is consistent with previous research on topic adaptation modeling through determination on whether a word falls within a given topic area or not (6). The document frequency results were expected to match the features reported by Boulis and Ostendorf, because they indicate the likelihood that the proportion of documents that contain a term differs significantly by gender, thus providing utility in document classification. The term frequency results were not expected to match up in the same fashion, as some terms that Boulis and Ostendorf reported, such as curse words, are relatively rare in speech in general—and words that are more common were expected to naturally dominate term frequency. The results are summarized in Figure 7.

| Document-based Chi-square | | | Feature-based Chi-square | | |
|---|---|---|---|---|---|
| Rank | Male | Female | Rank | Male | Female |
| 1 | hey | husband | 1 | uh | [laughter] |
| 2 | er | gosh | 2 | ah | mhm |
| 3 | wife | goodness | 3 | er | husband |
| 4 | shit | children | 4 | wordfragment | <sentence beginning> |
| 5 | pretty | husband's | 5 | yeah | <sentence ending> |
| 6 | man | daughter | 6 | you | oh |
| 7 | gotta | hi | 7 | mean | my |
| 8 | guy | son | 8 | [noise] | and |
| 9 | john | she | 9 | a | we |
| 10 | bucks | kids | 10 | wife | she |

Figure 7. Potentially discriminative features for feature recognition by gender.

As expected, the top document-based chi-square features were consistent with the types of words that are discriminative in gender classification, such as exclamations, names, and familial references. The one outlying element was "pretty", which our experiments showed was a top female-identifier in information gain on classification. In contrast, in the document chi-square "pretty" was shown to occur in far more male conversation sides than female.

The more interesting comparison exists between features most discriminative for gender classification and features which are most discernable from gender. With the exception of "husband" and "wife", the features identified from the feature-based chi-square are not content-bearing terms. Sentence delimiters are also indicated as female-predictable, possibly indicating that female speakers in this corpus are incurring more sentence breaks than male speakers. To test this, an analysis of our test set was performed: as predicted, female speakers averaged 9.5 words per sentence, while male speakers averaged 10.5 words per sentence, a significant difference. Topping male features are word fragments and empty speech markers such as "uh" and "er".

An analysis of the relative locations of the content-bearing words of the other lists such as "david" and "goodness" reveals that, although these words are lower on the feature-based chi-square index, they are still high up relative to other words on the list. Even in our most conservative interpolation strategy for reconciling gender-based and general language models, the features identified by Boulis and Ostendorf would still play a pivotal role.

The final experiments explored the question: can nonlinear interpolation methods on these data be used to reduce perplexity in automatic speech recognition systems?

Perplexity is a measure in information theory that is closely tied to entropy (log probability), which implies how "difficult" it is for a language model to choose the next word in an unknown system. Lower perplexities are generally indicative of more accurate language models, though this may not be true for small differences (<< 1%). Boulis and Ostendorf used bigram language models with Kneser-Ney smoothing. Our results were comparable with the original study, as shown in Figure 8.

| Wolf study | | |
|---|---|---|
| LM Used | On F | On M |
| F | 89.4 | 99.0 |
| M | 93.4 | 95.1 |
| Total Set | 88.1 | 93.7 |
| Same+Total | 87.9 | 93.3 |

| Boulis and Ostendorf study | | |
|---|---|---|
| LM Used | On F | On M |
| F | 82.8 | 94.2 |
| M | 86.0 | 90.6 |
| Total Set | 81.8 | 89.5 |
| Same+Total | 81.6 | 89.3 |

Figure 8. Perplexity of two-way bigram language models.

Figure 8 shows the perplexity characteristics of language models used on female-gender and male-gender test sets. The columns represent the two test sets, and the rows are the training models. Here, "total set" refers to the language model generated with male and female data combined, and "same + total" refers to the language model generated by interpolating the appropriate gender model to be tested (male on male, female on female) with the combined model, using a standard linear 50% interpolation of same-gender and combined language models through SRILM under Kneser-Ney discounting.

Several attempts were then made to improve the interpolated results. Initial tests attempted to show whether a modified interpolation weight would be sufficient to reduce perplexity by a significant margin. By adjusting the interpolation procedure, small but insignificant gains were made, with the greatest gain coming from a linear interpolation ratio of 70% from the total set, 30% from the gender model. This reduced perplexity by 0.2 points for both genders. Another set of tests examined the elimination of smaller counts in the group model from consideration in the interpolation, with unigrams and bigrams considered separately. Cutoff values for pruning unigrams and bigrams with fewer than X counts were tested, where X ranged between 2 and 5. In each case, gains were negligible.

Several attempts to modify the SRILM-created language models manually led to a rise in perplexity, likely because the backoff weights were not being recalculated correctly. Further tests examined the possibility of adjusting the word counts prior to language generation. By interpolating the word counts rather than the resulting probabilities, the language models would be consistent with the techniques used by SRILM to construct the language models. A mathematical interpolation on word counts is possible because Kneser-Ney smoothing allows for fractional counts in SRILM. Initial tests at modifying the word counts by dynamically increasing the counts listed for words based on the chi-square measure led to a dramatic rise in perplexity, however. This was in part because singleton counts were not being preserved, and the predictability of new words decreased accordingly. Also, potential confounding could exist between unigrams and bigrams, further skewing the prediction of new words. A general interpolation strategy that failed to account for singletons would not likely be viable.

To come up with a workable interpolation procedure, we focused on designing a simple system that would treat a chi-square cutoff region in the gender-based counts data as a tool for dividing the gender-based data into two groups of relevant and irrelevant features. The relevant feature counts would remain unchanged, preserving them for the interpolation procedure. The irrelevant feature counts would be reduced, but modified in such a way as to preserve singleton counts.

The formula to reduce weight for irrelevant counts was:

**modified value = ((observed value / 2) + 0.5)**

This formula was used as a baseline test to see if a ballpark perplexity figure could be achieved. It ensured that no singleton counts would be added to or subtracted from the resulting file, to eliminate the prior confound. Features scoring above a cutoff value (those less likely to be due solely to chance) would be considered "relevant counts" and remain unmodified, while features scoring below a cutoff value were considered "irrelevant counts" and were decreased using this equation.

A range of cutoff alpha values from 1:20 to 1:1,000,000 was tested for both feature-based and document-based chi-scores. Alpha values for a chi-square test are a probability translated directly from the chi-square statistic that a difference observed between the four categories (both genders crossed with both conditions of observed/not-observed) is due to chance. For example, a cutoff alpha value of 1:20 would mean that any features determined to have a 5% or greater probability of having the same frequency in both gender categories would be deemed irrelevant and reduced using the previous equation. Using this range of cutoff values was in keeping with the theory that discriminative features were the ones most likely to be truly different between genders.

The results of this approach were promising. While worse than interpolation without modification, they were still in the range of expected values for group data. For the female models, perplexity scores ranged from 89.3 to 89.5, as compared to the female model result of 89.4. For the male models, perplexity scores ranged from 94.9 to 95.0, as compared to the male model result of 95.1. None of the models produced through the interpolation with data modified by cutoff values were significantly different from each other.

# DISCUSSION

Although replication of the documentation classification experiments of Boulis and Ostendorf was largely consistent with their findings, several unexpected results occurred during the course of the experiments.

It is interesting to note that the performance in determining a speaker's gender based only on the transcript of the other speaker, while above chance for both studies, is significantly better in our study. One possible explanation is that inclusion of the individual word fragments added more information about when a given speaker is interrupted, and this information may be highly gender-dependent. This was reinforced by the discovery that word fragments were strongly correlated with male gender in our feature-based chi-square analysis.

The relative perplexity results were consistent with Boulis and Ostendorf's study, although in our study, the male gender data was better at predicting female data than male data. This was likely due to the fact that male speech was harder in general to predict, in both studies, thus limiting the comparability between the two testing sets.

The least expected result was the feature "pretty", which our experiments showed was a top female-identifier in information gain on the classification tests, but occurred in far more male conversation sides than female. Although the two tests used different randomizations of inherently uncertain conversational data, the results should be

impossible given the high chi-square value associated with document frequency of "pretty". Further analysis will be needed to isolate the problem.

Since the models interpolated from modified counts data were not significantly different, this suggests that the modifications being performed may not be drastic enough to sufficiently affect perplexity. A more dynamic approach to counts modification may provide a better opportunity for improving the models. However, it is notable that similar experiments with topic-based data which modified probabilities rather than counts have also turned up negligible gains over traditional interpolation methods (6).

Future work will need to consider a more dynamic nonlinear interpolation approach rather than a binary on/off-topic classification for individual features. Modifying the counts data is statistically simpler, but obfuscates the underlying work that SRILM performs in creating the language model. Manipulation of the probabilities is a more likely avenue of discovery. To simplify the process, probability manipulation may need to be limited to unigram counts until measurable gains can be demonstrated.

REFERENCES

(1) C. Boulis and M. Ostendorf. 2005. A Quantitative Analysis of Lexical Differences Between Genders in Telephone Conversations.

(2) A. McCallum. 2006. Bow: A toolkit for statistical language modeling, text retrieval, classification, and clustering. http://www.cs.cmu.edu/~mccallum/bow/

(3) C. Cieri, D. Miller, and K. Walker. 2004. The Fisher Corpus: a resource for the next generations of speech-to-text. In *4th International Conference on Language Resources and Evaluation, LREC*, pages 69–71.

(4) A. Stolcke. SRILM – An Extensible Language Modeling Toolkit. http://www.speech.sri.com/

(5) X. Huang, A. Acero, and H. Hon. 2001. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice-Hall, Inc. Upper Saddle River, NJ. 07458

(6) K. Seymore, C. Stanley, and R. Rosenfeld. 1998. Nonlinear Interpolation of Topic Models for Language Model Adaptation. *Proceedings of ICSLP98*, December.

(7) M. Koppel, Shlomo Argamon, and A. R. Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, v.17 n.4, pages 401-412.

(8) Author Argamon, S. Koppel, M. Fine, J. Shimoni, A. R. 2003. Gender, genre, and writing style in formal written texts. *Text, The Hague, Then Amsterdam, Then Berlin,* v.23 n.3, pages 321-346.

(9) Kiesling, Scott F. 2004. Dude. *American Speech*, v.79, n.3, Fall 2004, pp. 281-305.