Automatic Identification of Gender from Speech

Sarah Ita Levitan[1*], *Taniya Mishra*[2], *Srinivas Bangalore*[3]

[1]Dept. of Computer Science, Columbia University, New York, NY 10027, USA
[2]Interactions LLC, 25 Broadway, New York, NY 10004, USA
[3]Interactions LLC, 41 Spring Street, Murray Hill, NJ 07974, USA

sarahita@cs.columbia.edu, tmishra@interactions.com, sbangalore@interactions.com

Abstract

Identifying the gender of a speaker from speech has a variety of applications ranging from speech analytics to personalizing human-machine interactions. While gender identification in previous work has explored the use of the statistical properties of the speaker's pitch features, in this paper, we explore the impact of using spectral features in conjunction with pitch features on identifying gender. We present a novel approach that leverages pitch feature trajectories in the interest of identifying the speaker's gender with as little speech as possible. We also investigate the cross-lingual robustness of a model trained on English speakers to identify the gender of German speakers. Finally, we present a model for gender detection in German that outperforms the state-of-the-art results on a benchmark data set.
Index Terms: gender identification, human-computer interaction, computational paralinguistics, feature trajectories

1. Introduction

Automatic gender identification from speech is an important problem with many applications including speaker identification, speaker segmentation, and personalizing human-machine interactions. Knowledge of gender can be used for normalization of speech features, which has been shown to decrease word error rate [1] in speech recognition. Gender identification can improve the prediction of other speaker traits such as age and emotion, either by jointly modeling gender with age (or emotion) or in a pipelined manner. Speaker verification systems also implicitly or explicitly use gender information. In general, identification of a speaker gender is important for increasingly natural and personalized dialogue systems.

Previous work has examined the various differences between male and female speech. These differences include physiological (e.g. vocal tract length) [2], phonetic [3], and voice quality differences [4]. In studies of human perception of pitch location, absolute fundamental frequency (f0) has been found to be the most important information for deciding both relative pitch level and speaker gender. [5] Thus, f0 is a crucial feature for automatic gender identification.

There is rich literature on extracting para-linguistic information such as gender, age, dialect and emotion from speech in order to complement the words extracted using speech recognition. Early work on automatic identification of gender using speech features as input was developed in the late 1990s and early 2000s [6, 7, 8, 9] and the topic continues to be explored in recent work of [10, 11, 12].

In investigating the influence of gender on spoken utterances, past research has focused on prosodic features and cepstral features [13]. A typical approach to identifying gender (or other speaker traits) is to compute summary statistics of the speech features (pitch or spectral), such as mean, maximum, minimum, over a specific time span and to use these statistics as features in gender classification models. In this paper, we have modeled the relationship between gender and both prosodic and spectral feature sets, considering each feature set individually as well as jointly. We also introduce and investigate a novel approach that uses the entire feature trajectory for gender classification. This approach has the benefit of being low latency in the context of streaming speech input, and potentially provide a faithful representation of the input without the loss incurred in computing summary statistics.

In this paper, we compare and contrast the modeling choices when training a gender identifier using prosodic versus spectral feature streams, summary statistics versus complete feature trajectory and numerical versus categorical representation of input features. We also explore the cross-lingual robustness of our classifiers by evaluating the performance of our English-trained classifiers on German data. Finally, we investigate the impact of including child speech with adult speech and demonstrate an improved accuracy on a benchmark German data compared against previous results on this data.

2. Features For Gender Classification

2.1. Fundamental Frequency

We used Praat [14], an open-source toolkit for speech analysis, to extract the voice fundamental frequency (f0) trajectory of each utterance in the corpus. We extracted f0 trajectories using 10 ms, 5 ms, and 2 ms sampling, to examine whether more granular sampling adds useful information that can improve classification. For each of the trajectories, we computed the minimum, maximum, median, mean, and standard deviation of the f0 values. We used these summary statistics as features for our machine learning classifiers. In addition, we used the entire trajectory of f0 values as input, without any summarization for the MaxEnt-based classification approach (described below in Section 3.1.4).

2.2. Cepstral Coefficients

In addition to f0 statistics and trajectories, we explored the use of cepstral features for gender detection by using . We used Mel-frequency cepstral coefficients (MFCCs) as features. We generated 21 MFCCs and energy (dB scale) for each frame and computed the mean and standard deviation of each coefficient over a sliding window of 11 frames, centered on the current

*This work was done while the first author was an intern at Interactions LLC.

frame. This reduced the MFCC trajectory size and sped up performance.

3. Gender Detection Experimental Evaluation

3.1. Gender Detection in English

3.1.1. English Data

We used the HMIHY ("How May I Help You") corpus [8] for our gender detection experiments. This corpus has been used for gender detection tasks, enabling us to compare our approach and experimental results against previously published results. The HMIHY corpus was obtained from a natural language human-computer spoken dialogue system developed at AT&T, that enabled callers to speak with an automated agent. The speech data in this corpus are telephone conversations that sometimes include background speakers and noise.

The HMIHY corpus includes 5002 utterances from 1654 speakers, with an average utterance duration of about 6 seconds. We used a subset of 4520 utterances with gender labels for our experiments. Two human annotators labeled the corpus with perceived gender labels (male or female), with perfect agreement. We partitioned the corpus into 3 disjoint sets, with randomly selected utterances from distinct speakers in each subset: 80% training set, 10% development set, and 10% test set. The partitions were created by percentage of speakers, not number of utterances. Each split had 64% male speakers and 36% female speakers. The baseline (majority class) gender identification accuracy for this corpus is therefore 64%. The average f0 statistics for male and female speakers in the HMIHY training set is shown in Table 1.

Table 1: *Average f0 statistics by gender in the HMIHY training set.*

gender	minf0	maxf0	medianf0	meanf0	stdvf0
male	98.6	249.3	129.23	135.9	29.9
female	124.6	300.0	197.4	200.1	33.0

3.1.2. Rule-based Classification

From the statistics in Table 1, we saw that there is a substantial difference in f0 ranges between male and female speakers. In particular, we observed the largest gap is in the mean f0 between males and females. Leveraging this difference in the mean f0 between the two genders, we developed the following simple rule to serve as our baseline classifier: if the mean f0 of the speaker is closer to the average pitch of males than to the average pitch of females, we label the speaker as male. Otherwise, we label the speaker as female. We applied this simple classification approach to our entire test corpus, and achieved an accuracy of 87.2%. Although this rule relies on knowing the ground truth mean f0 for females and males, it indicates that a simple rule can fairly accurately predict gender from speech. In the rest of this paper, we attempt to improve the gender classification accuracy not only above the majority-class baseline (64%) but also above the meanf0 rule-based baseline (87%), by applying machine learning techniques to the gender detection task.

3.1.3. F0 and MFCC Statistics

Typically, the features used for gender detection and similar tasks are in the form of summary statistics of prosodic and spectral features, particularly in discriminative classification approaches. We trained 4 classifiers – logistic regression, linear regression, random forest and AdaBoost available from Python's scikit-learn library [15], on f0 summary statistics, MFCC summary statistics, and using a combination f0 and MFCC statistics. We experimented with different durations of the utterances to understand the relationship between duration and accuracy. We found that the logistic regression classifier performed the best for all three feature sets, therefore only those results are presented here.

Table 2: *F0 and MFCC statistics classification results of a logistic regression learner using varying length segments of speech.*

Duration	f0	MFCC	f0+MFCC
.5	90	67.8	91.3
1	89.4	77.2	91.1
1.5	90.9	82.4	90.1
2	92.4	87.8	95.2
2.5	93.3	89.1	94.8
2	93.5	91.3	94.4
all	93.3	92.8	95.2

Although we obtain the best gender identification accuracy using the entire utterance, with as little as 2 seconds of speech, as shown in Table 2, we obtain accuracies close to those obtained using the entire utterance.

Our results show that f0 features by themselves are more predictive of gender than cepstral features. We also find that f0 features are more useful immediately at the start of an utterance; we achieve 90% accuracy after only half a second of speech. In contrast, we observe a sharp increase in performance for the MFCC model as the amount of speech increases. This indicates that MFCC features (as we computed them over 11 frames) are useful over longer periods of speech.

The models using the combined f0 and MFCC features are consistently more accurate for all durations, at about 95% accuracy using 2 seconds of speech, providing credence for combining prosodic and spectral information.

We compare our results with those reported in [8] for the same task using the same training and test corpus. Although Shafran et al. did not train a model using f0 statistics alone, their HMM-based classifier trained on f0 and MFCC statistics achieves an accuracy of 95.4%. Our experiments show that we can obtain competitive results using only a short fragment of each utterance. Here, we obtain as high as 95% accuracy after only 2 seconds of speech, while Shafran et al. obtain their results using the entire utterance.

It is difficult to make further comparisons with other gender identification results because of differences in tasks and data. For example, [12] report gender identification accuracy as high as 98% using f0 features, however their data is cleanly recorded, and that performance might not be replicatable in a deployed system.

3.1.4. F0 Trajectories

Having established the state-of-the-art accuracies of our models trained on summary statistics-based features, we explore the

possibility of using the entire raw trajectory of f0 values for gender classification instead of summary statistics. This approach has the advantage of avoiding summary statistics computation in the context of streaming speech input where incremental, low latency, gender identification might be desired.

Unlike the summary statistics features which result in fixed dimension feature vectors for the classifier, using the f0 trajectories results in variable dimension feature vectors which most classifier implementations do not permit as inputs. Hence, we model the f0 trajectories as text input with each token of text corresponding to the binned f0 value to the nearest tens place. We used a maximum entropy (MaxEnt) text classifier, LLAMA [16] which computes up to trigrams as features on the input of a training example. We experimented with three f0 trajectories by extracting f0 every 10ms, 5ms, and 2ms.

We also trained a LLAMA model on binned f0 statistics, in order to directly compare results from categorial features on trajectories versus statistics. We binned the f0 statistics by rounding the minimum, maximum, mean, and median f0 values to the nearest tens place, and the standard deviation to the nearest tenths place.[1]

The results are shown in Table 3. When we compare the LLAMA trajectory approach with the LLAMA statistics approach, we find that at 10 ms sampling, the statistics approach outperforms the trajectory approach at every duration. However, consistent with our hypothesis, at 5 ms sampling, the trajectory approach outperforms the statistics approach LLAMA for durations 1, 1.5, 2, 2.5, and 3 seconds. At 2 ms sampling, the trajectory approach is better for 1.5 and 2 seconds. These results indicate that a trajectory approach for gender identification can be useful for applications where fairly accurate gender predictions are needed as quickly as 1 second. As far as we are aware, our trajectory approach to gender identification is a novel technique. Although the performance of this approach is lower than other machine learning models trained on f0 statistics, it is remarkable that viewing the f0 trajectory as text and modeling gender identification as a text classification problem can perform as well. It is possible that a more sophisticated quantization method of the f0 values can improve the performance and provide a simple and quick method to predicting gender.

Table 3: *F0 trajectory classification results.*

speech duration	f0 statistics	Sampling frequency 10 ms	5 ms	2 ms
0.5	90	86.5	89.1	88.9
1.0	89.4	88.0	**91.1**	89.4
1.5	90.0	89.8	**92.0**	**90.7**
2.0	90.7	90.2	**92.2**	**92.2**
2.5	92.6	92.0	92.0	92.4
3.0	93.5	90.7	91.3	90.9
all	93.3	91.7	92.4	90.2

3.2. Gender Detection in German

3.2.1. German Data

In this section, we report and compare the results of gender classification on a benchmarked data set, *aGender* corpus [17] of gender-labeled German speech utterances that has been used in the 2010 Interspeech Paralinguistic challenge. The challenge and the data are described in [18].

Subjects in the aGender corpus were given prompts to elicit specific types of speech, such as a date. The average utterance length in this corpus is 2.58 seconds. Each speech utterance in the corpus is labeled with one of three labels: *male*, *female*, and *child*. Speakers who are 14 years and under are marked as children. The training and development partitions in the aGender corpus have gender labels, however, the test set is unlabeled so we could not use this partition for our evaluation experiments. We instead report results from testing on the development set.

3.2.2. Male vs. Female vs. Children

The Interspeech gender challenge was a 3-way classification problem between male, female, and child speech. To compare our results with the benchmarked data, we trained models using f0 statistics and models where the f0 statistics features were supplemented with energy statistics such as min, max, median, and standard deviation of energy, as well as jitter and shimmer – features that represent voice quality. (We refer to this supplemented feature set as f0+.) We also combined these features with MFCC statistics. This is a more difficult task than the binary classification task of identifying *male* versus *female* problem. By including children, we introduce the challenge of differentiating between young childrens speech (which is normally high pitched) and high pitched female speech.

We present our results in Table 4. As a comparison to previous work, the organizers of the challenge provide a baseline system, trained on 450 acoustic features extracted using openSmile [19]. This baseline system achieved an accuracy of 76.99% . The best performing system submitted to the challenge [20] achieved an accuracy of 84.3%.

Table 4: *Gender identification with children.*

model	f0	f0+	MFCC,f0+
Logistic regression	81.2	82.2	84.9
Linear regression	81.8	81.8	-
Random forest	**83.3**	**84.1**	**85.0**
AdaBoost	83.2	84.1	82.7
LLAMA	82.2	82.1	-

All five of our f0-stats models perform better than the non-trivial baseline of 76.99%. Random forest yields the best model, achieving an accuracy of 85.0% when using all features, 83.3% with only f0 features, and 84.1% when using the f0+energy+voice quality feature set. The random forest model using all features achieves an accuracy of 85%, and is close to state-of-the-art for the 3-class gender identification problem. This performance is impressive, as the top-scoring system was based on a late fusion of six subsystems.[2]

3.2.3. Cross-lingual Gender Detection

The availability of the aGender corpus provides us with an opportunity to study the cross-lingual robustness of the previously described gender identification models that were trained on the

[1] These bins were determined by experimenting with development data.

[2] Note that we cannot make direct comparisons since the final results are on the test set, and we can only evaluate our results on the development set. However, the challenge organizers note that results from training on train+dev and testing on test data were usually higher than on the development data, due to the larger training data set.

HMIHY data. In order to utilize the HMIHY trained models that provides *male* and *female* labels, we excluded the data labeled as *children* from the German corpus. After excluding children and unvoiced utterances, we were left with the partitions shown in Table 5.

Table 5: *aGender partitions.*

split	utterances	percent female
train	28,051	50.3%
devel	18,106	53.2%

To explore the cross-lingual nature of spoken cues toward gender identification, we tested on the aGender data (18k utterances) 4 scikit-learn models and one LLAMA model trained on HMIHY data (3.6k utterances) using f0 summary statistics from 2 seconds of speech. Table 6 presents the performance of the models on the aGender corpus in the "Cross-lingual" column.

Table 6: *Cross-lingual and in-language gender identification results using f0 statistics.*

Learner	Cross-lingual	German only
Logistic regression	**92.1**	92.3
Linear regression	91.9	91.7
Random forest	92.0	**93.0**
AdaBoost	91.0	92.9
LLAMA	89.5	92.4

Our results show that training on English data and testing on German data yields performance well above the simple majority baseline (female; 53.24%) the highest accuracy of 92.1% is achieved using logistic regression on f0 summary statistics and are very similar to our results from training and testing on English HMIHY data. Using a logistic regression model for 2 seconds, we previously achieved an accuracy of 92.4%. This suggests that the pitch statistics capture crosslingual features about gender across languages. This finding is consistent with previous work on cross-language gender detection from speech [7], which tested an HMM-based gender identification system on 11 languages and found that gender detection is largely language independent.

Next, we compare the cross-lingual results against in-language gender detection with the same 2-classes, *male and female*, using the aGender data (train: 28k utterances, test: 18k utterances).

The results are shown in Table 6, in the "German-only" column. We find that the best accuracy, 93.0%, is achieved with a random forest classifier. We note that the LLAMA model performs better when trained and tested on the same language. It achieves an accuracy of 89.5% when training on English data and testing on German, and an accuracy of 92.4% when training and testing on the same language, German or English. This suggests that the LLAMA approach of binning and treating numeric features as strings is not robust across languages.

4. Conclusions and Future Work

We present the results of several experiments related to automatic gender identification from speech. In our first set of experiments, we find that using simple f0 summary statistics from only 2.5 seconds of speech, we can achieve the same results as reported in [8] using MFCC and f0 features on the entire utterance (about 6 seconds). We present a novel trajectory approach to gender identification, using the entire f0 trajectory of values as input to LLAMA, a categorical classifier. This achieves remarkable results considering that the numeric features are treated as ngrams, and suggests that the trajectory approach can be useful for obtaining fairly accurate gender predictions with as little as one second of speech. We compare MFCC and f0 features and find that f0 alone is more discriminative than MFCCs, but a combination of both feature streams yields the highest performance. We show that using MFCC and f0 features from only 2 seconds of speech, we can obtain the same results as [8] using similar features on the entire utterance.

In our cross-lingual experiments, we find that we can train a gender classifier on a relatively small English training set (3.6k utterances, 2 seconds each) and achieve a 92% accuracy testing on a large German corpus (18k utterances, 2 seconds each), almost as good as training and testing from the same German corpus (93.8%). Finally, our experiments with the 3-class problem male, female, child, result in a close to state-of-the-art system using a random forest model trained on simple f0 and MFCC features.

Future work using f0 trajectories can make use of a more sophisticated binning technique to better handle the numeric features. We can improve the 3-way classification by using additional child speech corpora for training. Additionally, the scikit-learn models in this work used the default parameters; tuning the parameters will likely result in increased performance. More broadly, the approaches in this work can be applied to many other paralinguistic detection problems such as age and emotion. In particular, it is possible that our trajectory approach can capture nuances in the f0 contour that summary statistics cannot, and can potentially improve performance of paralinguistic detection systems.

5. Acknowledgements

We thank Olivier Boeffard, Patrick Haffner, Ethan Selfridge, and Svetlana Stoyanchev for helpful conversations related to this work.

6. References

[1] S. Wegmann, D. McAllaster, J. Orloff, & B. Peskin, "Speaker normalization on conversational telephone speech", in *Proceedings of ICASSP 1996*, vol. 1, pp.339-341, 1996.

[2] D. R. Smith & R. D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgments of speaker size, sex, and age", in *The Journal of the Acoustical Society of America*, vol. 118-5, pp.3177-3186, 2005.

[3] A. P. Simpson, "Phonetic differences between male and female speech", in *Language and Linguistics Compass*, vol. 3-2, pp.621-640, 2009.

[4] C. G. Henton, "Fact and fiction in the description of male and female pitch", in *Language and Communication*, vol. 9, pp.299-311, 1989.

[5] J. Bishop, & P. Keating, "Perception of pitch location within a speaker's range: Fundamental Frequency, voice quality and speaker sex", in *The Journal of the Acoustical Society of America*, vol. 132-2, pp.1100-1112, 2012.

[6] R. Vergin, A. Farhat, & D. O'Shaughnessy, "Robust gender-dependent acoustic-phonetic modeling in continous speech recognition based on a new automatic male/female classification", in *Spoken Language*, vol. 2, pp.1081-1084, 1996.

[7] E. S. Parris, & M. J. Carey, "Language independent gender identification", in *Proceedings of ICASSP 1996*, vol. 2, pp.685-688, 1996.

[8] I. Shafran, M. Riley, and M. Mohri, "Voice signatures", in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 31–36, 2003.

[9] H. Harb, and L. Chen, "Gender identification using a general audio classifier", in *Proceedings of ICME 2003*, pp. 733–736, 2003.

[10] M. Li, K. Han, J. Kyu, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion", in *Journal of Computer Speech & Language*, vol. 27-1, pp. 151–167, 2013.

[11] S. Safavi, P. Jancovic, M. J. Russell, and M. J. Carey, "Identification of gender from children's speech by computers and humans.", in *Proceedings of Interspeech 2013*, Lyon, France, pp. 2440–2444, 2013.

[12] Y. Hu, D. Wu, & A. Nucci, "Pitch-based gender identification with two-stage classification.", in *Security and Communications Networks*, vol. 5, pp.211–225, 2012.

[13] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, & S. Bennett, "Robust prosodic features for speaker identification.", in *Proceedings of ICSLP 1996*, vol. 3, pp.1800–1803, 1996.

[14] P. Boersma, and D. Weenink, "Praat, a system for doing phonetics by computer .", in *Glot International*, pp.341-345, 2001.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, ... & E. Duchenesay, "Scikit-learn: Machine learning in Python.", in *The Journal of Machine Learning Research*, vol. 12, pp.2825–2830, 2011.

[16] P. Haffner, "Scaling large margin classifiers for spoken language understanding.", in *Speech Communication*, vol. 48, pp.239–261, 2006.

[17] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech.", in *Proceedings of LREC 2010*, Valletta, Malta, pp.1562–1565, 2010.

[18] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language–state-of-the-art and the challenge.", in *Computer Speech & Language*, vol. 1, pp.46–64, 2013.

[19] F. Eyben, F. Weninger, F. Gross, & B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor.", in *Proceedings of ACM Multimedia*, Barcelona, Spain, pp.835-838, 2013.

[20] H. Meinedo, & I. Trancoso, "Age and gender classification using late fusion of acoustic and prosodic features", in *Proceedings of Interspeech 2010*, Makuhari, Japan, pp.2818-2821, 2010.