Gender Identification by Voice

Kunyu Chen

Stanford University
kunyu@stanford.edu

I. Introduction

Gender identification by voice is useful in speech-based recognition systems which employ gender-dependent models. Vogt and André [1] suggested that gender differentiation help improve automatic emotion recognition from speech. Harb and Chen [2] reported that classifying speaker's gender is an important task in the context of multimedia indexing. Our paper will examine the applicability of standard machine learning techniques to the voice-based gender identification problem.

II. Dataset

We use a subset of TIMIT Acoustic-Phonetic Continuous Speech Corpus, which is publicly available online. The corpus consists of 160 sentence recordings by 8 female and 8 male speakers. We only use 127 recordings because the rest have serious noise. The recordings are in wav format and the sampling rate is 16 kHz.

On Page 4, Figure 1 plots the amplitude of audio waves over time of a sentence recording of a female speaker. Figure 2, on Page 4, plots the recording of the same sentence while the speaker is a male.

The complete TIMIT Acoustic-Phonetic Continuous Speech Corpus contains recordings of 630 speakers, and each speaker reads 10 sentences. Some of the sentences are shared among speakers, and Figure 1 and Figure 2 plots a sentence that is spoken by them all.

III. Features and Preprocessing

We utilize Yaafe to extract audio features out of sentence recordings. Our strategy is to first build models with all the features, 24 in total, that Yaafe is able to extract. We then evaluate the performance of different models and run backward search for feature selection on the most performant model.

Yaafe takes *blockSize* and *stepSize* as input parameters for all available features. *blockSize* defines the frame size, the width of a sliding window over which Yaafe computes feature values. *stepSize* is the step between consecutive frames. The first frame is always centered on the first signal sample, with *blockSize/2* 0ś padded to the left. Whenever the number of signal samples is not enough for the last frame to have *blockSize* samples, 0ś are padded to the right. For illustration, if *blockSize*=8 and *blockSize*=4, the first frame is centered on the 1st signal sample and the second frame is centered on the 5th. Therefore, the first frame has 4 padded zeros together with the first 4 signal samples. While the second frame covers the first 8 signal samples.

We use *blockSize=1024* and *stepSize=512* for all features. Each feature's frame coincides perfectly, therefore the set of all features at each frame can be treated as a single data point. We have 12004 such data points, among which 6286 are labeled 'female' and 5718 are labeled 'male'. We randomly pick up 8402 data points (70%) as training data, and the rest serves as test data for cross validation.

IV. Models and Results

We train naive Bayes (NB), discriminant analysis (DA), support vector machine (SVM) with linear kernel, nearest neighbor (NN) and classification tree (CT) classifiers with the training data , and we test the models against the test set. Table 1 on page 4 summarizes the results.

Note, when fed with all available features,

the discriminant analysis (DA) classifier is most performant in terms of test error rate and precision. We do not include linear regression nor generalized linear models here because with all available features, model terms are rank deficient.

We run backward search for feature selection on the discriminant analysis classifier. The "Test Error Rate" column of Table 2 on Page 4 presents the performance measure when we start with 24 features and iteratively remove one feature from the model at a time. The feature is selected so that the new model with one fewer features has the minimum test error rate. We apply similar step-wise greedy algorithms to get the other two columns.

Observe that the discriminant analysis classifier with 4 features performs even better than the model with 24 features.

V. Discussion

Our best performant model still suffers from a test error rate of greater than 10%. In order to better understand the nature of our classification problem, as well as to direct our future research in the right direction, we run diagnostics to see if our model has high variance or high bias. Figure 3 and Figure 4 on Page 4 are the learning curves of two discriminant analysis models, one with 4 features and the other one with all available features.

It turns out that even the all-feature discriminant analysis model depicts a typical learning curve for high bias. The high bias problem implies that the set of all available features we are considering does not capture enough gender-specific characteristics of voice.

VI. Conclusions

Our experiments involve applying standard machine learning techniques to the voice-based gender identification problem. Discriminant analysis works well and we are abel to achieve 88% accuracy, precision and recall. By running backward search for feature selection and diagnostics, we better understand the structure of the problem. In addition, we also conclude that general-purpose audio features may not be able to capture enough gender-specific characteristics of voice.

VII. Future

Zeng, Wu, Falk, and Chan [6] reported that applying Gaussian Mixture Models combined with high order audio features as parameters achieves robust results. It would be interesting to introduce high order audio features to our models and see its impact on our high bias problem. In addition, Harb and Chen [7] also published promising results using neural networks.

References

[1] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*. Citeseer, 2006.

[2] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," *Journal of Intelligent Information Systems*, vol. 24, no. 2-3, pp. 179–198, 2005. [Online]. Available: http://dx.doi.org/10.1007/s10844-005-0322-8

[3] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," http://web.mit.edu/6.863/share/nltk_lite/timit/, 1993.

[4] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, August 9-13 2010, pp. 441–446, http://ismir2010.ismir.net/proceedings/ismir2010-75.pdf.

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A li-

brary for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[6] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, "Robust gmm based gender classification using pitch and rasta-plp parameters of speech," in *Machine Learning and Cybernetics, 2006 International Conference on*. IEEE, 2006, pp. 3376–3379.

[7] H. Harb and L. Chen, "Gender identification using a general audio classifier," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 2, July 2003, pp. II–733–6 vol.2.

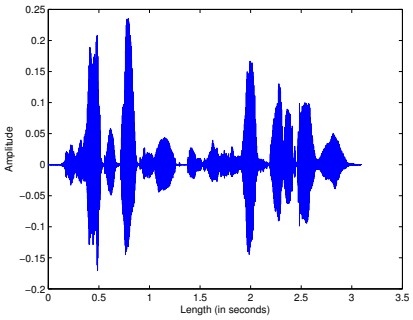

Figure 1: *Voice of a female speaker*

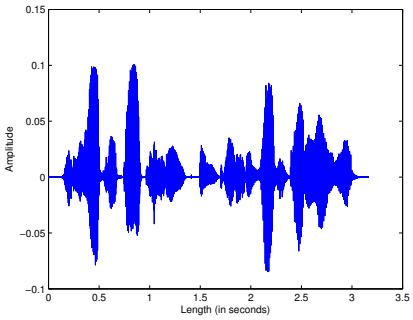

Figure 2: *Voice of a male speaker*

	Test Error Rate	Precision	Recall
NB	45.59%	75.00%	19.41%
DA	13.83%	86.88%	86.69%
SVM	35.92%	59.38%	99.36%
NN	42.73%	59.39%	58.17%
CT	16.88%	84.80%	82.56%

Table 1: *Performance of different models built with all available features*

# of Features	Test Error Rate	Precision	Recall
1	17.91%	77.81%	100.00%
2	15.32%	84.41%	91.46%
3	14.10%	86.10%	90.24%
4	12.77%	87.57%	88.92%
8	12.10%	88.28%	88.92%

Table 2: *Performance of DA models built with a select subset of features*

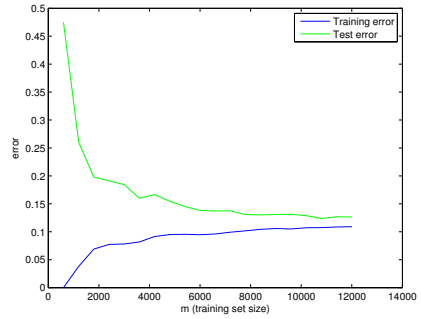

Figure 3: *Learning curve for a 4-feature DA model*

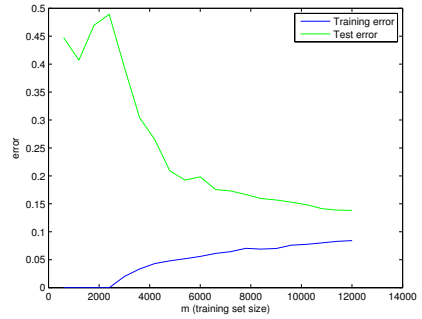

Figure 4: *Learning curve for an all-feature DA model*