# A comparative study of automatic classifiers to recognize speakers based on fricatives

*Nina Hosseini-Kivanani*[1], *Homa Asadi*[2&3], *Christoph Schommer*[1], *Volker Dellwo*[3]
[1] University of Luxembourg, [2] University of Isfahan, [3] University of Zurich

Nina.hosseinikivanani@uni.lu, h.asadi@fgn.ui.ac.ir, Christoph.schommer@uni.lu, volker.dellwo@uzh.ch

**Abstract (max 300 words)**

Speakers' voices are highly individual and for this reason speakers can be identified based on their voice. Nevertheless, voices are often more variable within the same speaker than they are between speakers, which makes it difficult for humans and machines to differentiate between speakers (Hansen, J. H., & Hasan, T., 2015). To date, various machine learning methods have been developed to recognize speakers based on the acoustic characteristics of their speech; however, not all of them have proven equally effective in speaker identification, and depending on the obtained techniques, the system achieves a different result. Here, different machine learning classifiers have been applied to identify the best classification model (i.e., Naïve Bayes (NB), support vector machines (SVM), random forests (RF), & *k*-nearest neighbors (KNN)) for categorizing 4 speaking styles based on the segment types (voiceless fricatives) considering acoustic features of center of gravity, standard deviation, and skewness. We used a dataset consisting of speech samples from 7 native Persian subjects speaking in 4 different speaking styles: read, spontaneous, clear, and child-directed speech. The results revealed that the best performing model to predict the speakers based on the segment type was RF model with an accuracy of 81,3%, followed by SVM (76.3%), NB (75.4%), and KNN (74%) (Table 1). Our results showed that the RF performed the best for voiceless fricatives /f/, /s/, and / ʃ / which may indicate that these segments are much more speaker-specific than others (Gordon et al., 2002), and the model performance was low for the voiceless fricatives of /h/ and /x/. Performance can be seen in the confusion matrix (Figure 1), which produced high precision and recall values (above 80%) for /f/, /s/ and / ʃ / (Table 2). We found that the model performance improved when the data related to clear speaking style; the information in individual speakers (i.e., voiceless fricatives) are more distinguishable in clear style than other styles (Table 1).

[Abstract Word count: 300]

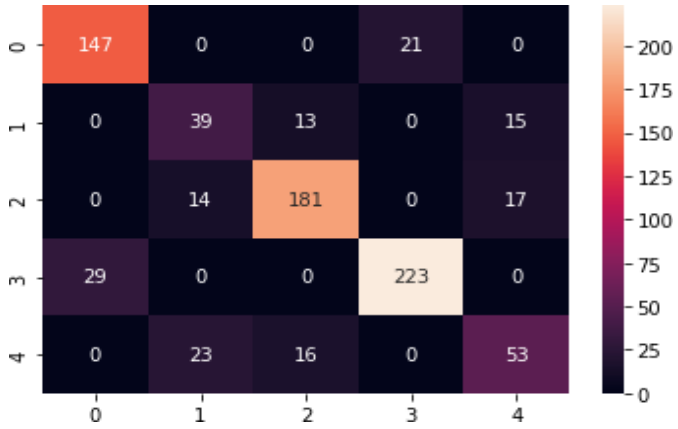


*Figure 1*: *Confusion Matrix of RF (0: f, 1: h, 2: s, 3: ʃ, 4: x)*

*Table1*: The output of ML models per speaking styles.

| Speaking styles | NB | SVM | RF | KNN |
|---|---|---|---|---|
| Read | 70% | 72% | 75% | 72% |
| Spontaneous | 61% | 63% | 65% | 65% |
| clear | **75%** | **76%** | **81%** | **74%** |
| Child-directed speech | 60% | 62% | 67% | 62% |

*Table2*: The output of RF model per segment types: clear speaking style

| RF | Precision | Recall | F1-score |
|---|---|---|---|
| /f/ | 84% | 83% | 83% |
| /h/ | **46%** | **55%** | **50%** |
| /s/ | **93%** | **84%** | **88%** |
| / ʃ / | 88% | 89% | 88% |
| /x/ | **58%** | **63%** | **60%** |

## References

Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, *32*(6), 74-99.

Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, *32*(2), 141-174.