# VOICE COMPARISON ANALYSIS OF FORENSIC RECORDINGS USING THE VOICESAUCE PROGRAM

Michael Jessen, Canan Konrat and Jonathan Horn

Department of Text, Speech and Audio, Bundeskriminalamt, Germany
{michael.jessen | canan.konrat | jonathan.horn}@bka.bund.de

## ABSTRACT

The computer program VoiceSauce was applied to a set of authentic telephone-based forensic recordings in order to examine the speaker-discriminatory potential of the acoustic voice quality information extracted with the program. Several conditions were tested, including variation on the input audio (all-vocalic vs. /a/-based) and the option of uploading hand-corrected formant and f0 streams. Among the voice quality parameters investigated, CPP (cepstral peak prominence) turned out to be the most speaker-distinguishing one, yet it stayed behind formant frequencies when used as baseline parameter. Judged from the performance indices EER and Cllr, speaker discrimination for the forensic data analysed with VoiceSauce remained quite limited. Multivariate models with several voice quality parameters, as well as logistic regression fusion, did not improve performance over the use of single parameters.

**Keywords**: forensic voice comparison, laryngeal voice quality, VoiceSauce

## 1. INTRODUCTION

Voice quality is a feature complex widely used in forensic phonetics [1]. Most commonly, forensic voice quality is judged auditorily [2, 3]. The focus of this paper is on laryngeal voice quality, also known as phonation type, as opposed to supralaryngeal voice quality also referred to as articulatory settings; henceforth "voice quality" will be used in the narrow sense of phonation type. Due to limitations in the intra- and interrater reliability of auditory voice quality ratings [4], being able to measure voice quality acoustically would be an advantage. There has been scepticism, however, that most acoustic voice quality measurements might not very well survive the telephone-passband or other forms of signal degradation typical in forensics [2]. Confirming such reservations, Enzinger et al. [5], using software called Glottex, found a speaker discrimination reduction in telephone-transmitted speech as compared to high-quality speech. However, Hughes et al. [6], using VoiceSauce [7], presented more mixed results, and especially with harmonic amplitude parameters there was little difference between studio and telephone quality, even when signal quality was further reduced within the latter category.

The main goal of this study is to explore further to which extent the program VoiceSauce can provide speaker-discriminatory information when applied to forensically realistic recordings. Though the focus of this study is on this program, there is an interest in forensic-phonetic acoustic voice quality research more generally.

## 2. METHOD

### 2.1. Forensic recordings

This study is based on a speech corpus called GFS 2.0 (German Forensic Speech). The corpus contains two recordings each of 23 male adult speakers of German. These recordings derive from anonymized authentic forensic cases involving telephone conversations (converted into PCM wav files sampled at 16 bit, 8 kHz). Single recordings from additional 25 speakers were used for the UBM (Universal Background Model), addressed in section 2.4. This corpus was also used in [8, 9]; the specific version applied is the one of [9], in which one of the 23 speakers was not included due to difficulties in measuring formant structure. Signal quality is on average lower than in non-forensic telephone recordings. In terms speech style, sections of speech with increased vocal effort or emotional involvement are included, but severe levels of signal distortion or non-neutral speech style were avoided in the compilation of the corpus. Net duration of the recordings is between about 20 and 60 seconds.

### 2.2. Audio data preparation

The measurements with VoiceSauce are based on vowel-only speech material. In some of the voice comparison tests (test condition design summarized in Table 1) the material used was the same as in [9], where for the purpose of long-term formant analysis (LTF) each recording was cut down manually to a 10-seconds long stream that contained only vowels with visible F1, F2 and F3. This condition will be called "LTF".

In other tests only the vowel /a/ was included in the stream. Depending on the number and lengths of

a-tokens in the recordings, the duration of the a-stream differs between recordings from between about 1.5 to 4 seconds. Tokens of /a/, regardless of phonological context, were identified based on listening from roughly sentence-level context, but when listening to the selected tokens in isolation, they had a broader range of mid-to-low vowel qualities. This condition is referred to as "a-tokens".

In both the LTF- and the a-token-data, cutting passages out of the recordings resulted in abrupt transitions between the remaining segments. In the spectrogram, these cuts may be visible as very short transients across the frequency range. During the brief moments of these transients, spectral tilt is radically reduced. Since spectral tilt is among the parameters measured with VoiceSauce, results could be affected negatively by the existence of these transients. To investigate that possibility, a subset of the a-token data was processed in a way that hard cuts were eliminated. This was performed by writing a Praat script that carried out the following steps for each of the a-tokens audio files: The a-tokens had been labelled in Praat. For each of the labelled tokens per audio file, the first zero crossing was identified and the signal from the beginning of the labelled segment to that zero crossing was muted. The same procedure was applied in the right-to-left direction of the segment. Subsequently all the segments per file were concatenated. The condition using data processed in this manner is called "a-tokens-zero".

### 2.3. VoiceSauce analysis

The program VoiceSauce (accessed August 8, 2022 from www.phonetics.ucla.edu/voicesauce) with compiled Matlab executables for Windows was used to analyse the audio data described in 2.2. Out of the voice source parameters offered by the program, the following set was used for analysis (see the program manual for their definition): H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-H2K*, H2K*-H5K, CPP, HNR05, HNR15, HNR25, HNR35. For reasons explained below, the parameter H2K* was added as well. VoiceSauce requires good estimation of formant frequencies and fundamental frequency in order to find the spectral events necessary for measurement of the voice source parameters. The Snack algorithm was selected for formant analysis and Praat for f0-analysis. Snack was selected due to compatibility with Wavesurfer [10], used in the LTF study in [9]. Praat was preferred over other f0-algorithms based on relative success (frequency and severity of f0 outlier values as criterion) on a subset of the data. Parameter defaults were used, except for frame shift, which was set to ten milliseconds instead of one, and Max F0 in Praat, which was set to 250 Hz.

Instead of automatic f0 and formant estimation, VoiceSauce also offers the option of manual data input. This option was used for some of the voice comparison test conditions. It allows the user to upload files containing hand-measured or manually corrected formant- or f0-value streams into the software, and in that case the uploaded values will be used for subsequent voice quality parameter extraction instead of the automatically measured f0 and formant values. In some test conditions only formants were uploaded. The formant values were taken from the formant measurements of [9] performed with Wavesurfer. Alignment across VoiceSauce parameters was ensured by taking the automatically calculated formant values and temporally aligning them with the uploaded formant values, after which only the uploaded values were used in the analysis. Maximum alignment was achieved by using a value of 2 for the data offset parameter in VoiceSauce. F1 was rarely in need of correction, therefore formant upload was limited to F2 and F3. In another test condition both hand-corrected formants and f0-values were uploaded. Uploaded f0-values were measured and corrected in Wavesurfer. Alignment was again achieved with a data offset value of 2. Pre-analysis showed that VoiceSauce-measured formants (F2 and F3) were more error prone with the data used here than f0. This is the reason why only in one test condition f0 data were uploaded in addition to formant data.

Table 1 shows the test conditions of this study. The term test means that questioned-speaker recordings from 22 speakers are compared against suspect recordings from the same 22 speakers, yielding 462 different-speaker comparisons and 22 same-speaker comparisons. Within a test condition the test is repeated several times while varying the voice quality parameter and the use or non-use of MAP, as well as two values for the number of Gaussians, all explained in 2.4.

| Cond. | Audio | Upload (manual data input) |
|---|---|---|
| 1 | LTF | - |
| 2 | LTF | F2, F3 |
| 3 | LTF | F2, F3, f0 |
| 4 | a-tokens | - |
| 5 | a-tokens | F2, F3 |
| 6 | a-tokens-zero | - |

**Table 1**: Test conditions used in this study.

### 2.4. Forensic-phonetic analysis

For each test as defined in 2.3, each of the 484 comparisons resulted in a score by using the GMM-

UBM approach. This approach was originally developed for automatic speaker recognition [11] and was adapted to the analysis of acoustic-phonetic data by Becker et al. [12]. In the current study the software VOCALISE version 1.6 [13] was used. The software accepts the frame-by-frame output generated by VoiceSauce for each of the voice quality parameters. This stream of values is modelled as a Gaussian Mixture Model (GMM) for the suspect recording and the 25-speaker set representing the UBM. The questioned-speaker values are not modelled in that manner, but instead they are evaluated frame-by-frame against the GMM of the suspect and the UBM, technically resulting in a Likelihood Ratio. These results are averaged across the frames. Further details about GMM-UBM analysis as applied here are the same as in [9].

In the tests, one of the parameters varied is the number of Gaussians for the GMM. In [9] the value of 17 was used. In pre-analyses it was tested whether this is a viable value for the current data. This value was accepted and a lower value of 3 was selected as well, because analysis of acoustic-phonetic data, formants in particular, tend to saturate at lower values of Gaussians as far as speaker-discrimination performance is concerned [12, 14]. The other parameter varied is the inclusion or exclusion of MAP (Maximum a posteriori) adaptation [9, 11]. One of the goals of MAP adaptation is to provide a more stable suspect model in case of insufficient suspect data. This might be relevant here with the a-tokens in particular.

For each individual voice quality parameter speaker discrimination performance was measured with EER (Equal Error Rate, using the convex hull method) and Cllr (log-likelihood-ratio cost) [15]. For the comparisons reported with Cllr, prior calibration was performed using cross validation logistic regression calibration [16].

Further analyses were made by grouping several parameters. In combination tests, the parameters of the group were included in a multivariate model, which is a regular option with the GMM-UBM approach. In fusion tests, the parameters were fused using logistic regression fusion [16]. Three sets of parameters were grouped for this purpose. The first group consists of all parameters using relative harmonic amplitudes (except, H2K*-H5K, which, as will be explained, is replaced by H2K*), the second one consists of those targeting spectral noise (CPP and HNR), and the third group consists of all parameters contained in the previous two groups. For the combination results Cllr is reported after calibration, for the fusion results Cllr is reported directly because fusion has a calibrating effect. For the functions described in this paragraph and the preceding one, the software BIO-METRICS was used [17].

## 3. RESULTS

Results are reported in two ways. First, numerical detail is presented for Condition 2. Secondly, in a more descriptive manner, patterns are pointed out that are similar or different between Condition 2 and the other five ones. A full report of all numerical results across conditions is made available at www.researchgate.net/profile/Michael-Jessen-2.

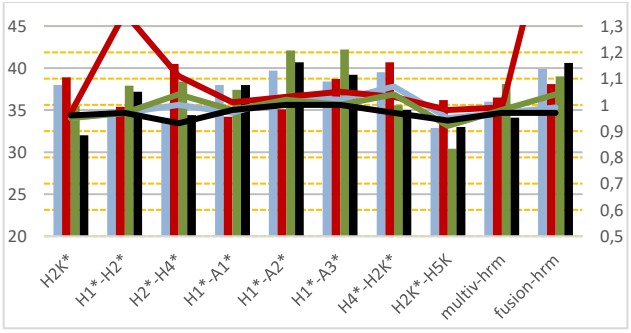


**Figure 1**: EER (bars; left vertical axis) and Cllr (lines; right) in Condition 2 for harmonic amplitude parameters.

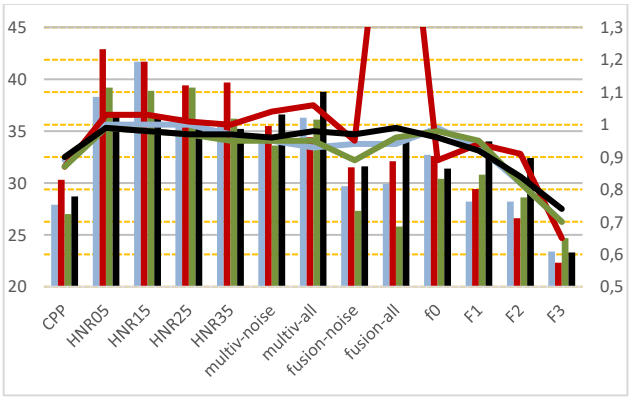


**Figure 2**: EER and Cllr in Condition 2 for spectral noise parameters, combination and fusion across all voice quality parameters, f0 and formant frequencies F1, F2 and F3.

For test condition 2 (see Table 1), Figs. 1 and 2 show the EER results (bars) and the Cllr results (lines) in the following four test variations (bars from left to right; same colours/greyscales for lines): no MAP, 3 Gaussians; no MAP, 17 Gaussians; MAP, 3 Gaussians; MAP, 17 Gaussians. Lower values indicate better performance. EER (in percent) of 50 and Cllr of 1 are at chance level. Looking at EER, the four test variations have an impact on the results, but there are no systematic performance differences. There are some differences between the voice quality parameters however. Among the harmonic amplitude parameters (Fig. 1), the best performance of single

parameters is with H2K*-H5K. Paradoxically, this is a parameter that should bear little meaning since the sampling rate of the recordings was 8 kHz and therefore 5 kHz is not represented. Could this result be driven by the component H2K*? In order to investigate this question, this parameter was included. Fig. 2 indeed shows that H2K*, although it is not a relative amplitude measurement (subtraction), is among the better-performing parameters, but it is not as good as H2K*-H5K (although in Conds. 4 and 6 it is slightly better). This has to remain an open issue. Applying multivariate processing or fusion to all harmonic amplitude parameters (excluding H2K*-H5K due to its questioned validity) leads to no particular performance advantage over single parameters. Among the spectral noise parameters (included in Fig. 2), a striking result is that CPP has the best performance. It is surpassed only by F3, which along with f0, F1 and F2 was included for comparison purposes with more familiar forensic-phonetic parameters as baseline. Fusing spectral noise parameters or applying a multivariate model did not improve upon CPP. The same holds for fusion and multivariate models across all parameters (but H2K*-H5K), except for fusion in one of the four variations. That particular fusion advantage, however, is not paralleled by Cllr. There was also an exception with Cond. 3, where noise-based or overall fusion slightly improved upon CPP.

Turning to the Cllr results, they roughly mirror the EER results. For example, Cllr is relatively low with CPP and still lower with F3, just like EER. There are some differences in detail however. In a few cases Cllr is massively out of scale. This is due to outliers in the scores, which Cllr is sensitive to but EER is not. The outliers occur in one of the two variations that make no use of MAP. MAP has the effect of preventing such extreme scores. Outlier values that occur without MAP were further negatively enhanced by cross validation calibration.

When comparing the voice quality results from Cond. 2 to the results from the other conditions mentioned in Table 1 (not shown, but provided through link), no major differences occur. It is not the case that single methodological steps such as uploading formants as opposed to fully automatic processing or removal of transients left any clear marks on the results. Most remarkable is the lack of any major difference between the conditions using 10-seconds LTF data and those using much shorter /a/-vowel data. Although there is a tendency for results to be better in the LTF data, this effect is quite small. Absence of a clear difference could mean that the extraction of the relevant information can be achieved with very short vocalic material, or it could mean that there is a trading relation in which the advantage for voice quality analysis of limiting the analysed material to the mid-to-low range of the vowel space [cf. 18, 19] can be traded against the advantage of increased duration of the vocalic input. Interestingly, formants (F2 and F3) suffer more in performance when switching from LTF to a-vowels than voice quality (shown in the link-provided data).

## 4. DISCUSSION

Speaker discrimination studies with VoiceSauce or related methods have been conducted predominantly on the basis of microphone-quality speech. Examples of such studies, with promising results, include [20-23]. There is less work on telephone-based or otherwise signal-degraded speech. A previous study that is of particular relevance to the present research is Hughes et al. [6].

VoiceSauce measurements on the forensic speech corpus GFS 2.0 have resulted in speaker discrimination performance that is quite limited. EER most of the time turned out to be in the range of 30 to 40% and only the parameter CPP showed better performance, reaching down into the 25-30% range. This is close to the performance of some formants based on the same material, though F3 has better performance still. Cllr was never below 0.8 even with CPP, and sometimes it was around chance level. Hughes et al. [6] in their telephone conditions differing in landline/mobile status and quality reported better values. For their multivariate models that correspond closely to “multiv-noise” and “multiv-hrm” in Figs. 1 and 2, they on average reported EER of about 20% and Cllr of between 0.7 and 0.9 for the former and for the latter EER of about 15% and Cllr slightly below 0.6. This is better performance than found here, especially for the harmonic amplitude parameters, which is contrary to the present results, where additive noise was better-performing than harmonic amplitude (probably CPP dominated in the multivariate model, so it really narrows down to that single additive noise parameter). Possible reasons for the performance difference between [6] and this study, respectively, include: telephone-transmission mostly simulated from microphone data vs. original telephone-recordings; fixed vs. varying signal quality across recordings; 60 seconds vs. 10 seconds vowels or less (a-tokens); contemporaneous speech vs. non-contemporaneous speech.

With more in-depth scrutiny of the extraction process it could be investigated further whether the limitations observed in this study are intrinsic to the forensic audio material or whether they result from difficulties of the algorithms to detect the relevant spectral events.

## 5. REFERENCES


[1] Gold, E., French, P. 2011. International practices in forensic speaker comparison. *Int. J. Speech, Lang. Law* 18, 293–307.

[2] Nolan, F. 2005. Forensic speaker identification and the phonetic description of voice quality. In: Hardcastle, W. J., Mackenzie Beck, J. (eds), *A Figure of Speech. A Festschrift for John Laver*. Lawrence Erlbaum Associates, 385–411.

[3] San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., Kavanagh, C. 2019. The use of the vocal profile analysis for speaker characterization: Methodological proposals. *J. Int. Phon. Assoc.* 49, 353– 380.

[4] Kreiman, J., Gerratt, B. R., Ito, M. 2007. When and why listeners disagree in voice quality assessment tasks. *J. Acoust. Soc. Am.* 122, 2354–2364.

[5] Enzinger, E., Zhang, C., Morrison, G. S. 2012. Voice source features for forensic voice comparison – an evaluation of the GLOTTEX software package. *Proc. Odyssey* Singapore, 78–85.

[6] Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., Gully, A. J. 2019. Forensic voice comparison using long-term acoustic measures of laryngeal voice quality. *Proc. 19th ICPhS* Melbourne, 1455–1459.

[7] VOICESAUCE software. http://www.phonetics.ucla.edu/voicesauce/

[8] Solewicz, Y.A., Jessen, M., van der Vloed, D. 2017. Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition. *Proc. Interspeech* Stockholm, 2849–2853.

[9] Jessen, M. 2021. MAP adaptation characteristics in forensic long-term formant analysis. *Proc. Interspeech* Brno, 411–415.

[10] WAVESURFER software. https://www.speech.kth.se/wavesurfer/index2.html

[11] Reynolds, D., Quatieri, T., Dunn, R. 2000. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* 10, 19–41.

[12] Becker, T., Jessen, M., Grigoras, C. 2008. Forensic speaker verification using formant features and Gaussian mixture models. *Proc. Interspeech* Brisbane, 1505–1508.

[13] VOCALISE software. https://oxfordwaveresearch.com/products/vocalise/

[14] Jessen, M., Alexander, A., Forth, O. 2014. Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. *Proc. Audio Engineering Soc. 54th Int. Conf. Audio Forensics* London, 28–35.

[15] Van Leeuwen, D. A., Brümmer, N. 2007. An introduction to application-independent evaluation of speaker recognition systems. In: Müller, C. (ed), *Speaker Classification I: Fundamentals, Features, and Methods*. Springer, 330–353.

[16] Morrison, G. S. 2013. Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Aust. J. Forensic Sci.* 45, 173–197.

[17] BIO-METRICS softw. https://oxfordwaveresearch.com/products/bio-metrics/

[18] Hanson, H. M. 1997. Glottal characteristics of female speakers: Acoustic correlates. *J. Acoust. Soc. Am.* 101, 466–481.

[19] Hanson, H. M., Chuang, E. S. 1999. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *J. Acoust. Soc. Am.* 106, 1064–1077.

[20] Jessen, M. 1997. Speaker-specific information in voice quality parameters. *Forensic Linguistics* 4, 84–103.

[21] Vaňková, J., Skarnitzl, R. 2014. Within- and between-speaker variability of parameters expressing short-term voice quality. *Proc. Speech Prosody* 7, 1081–1085.

[22] Kreiman, J., Park, S. J., Keating P. A., Alwan, A. 2015. The relationship between acoustic and perceived intraspeaker variability in voice quality. *Proc. Interspeech* Dresden, 2357–2360.

[23] Lee, Y., Keating, P. A., Kreiman, J. 2019. Acoustic voice variation within and between speakers. *J. Acoust. Soc. Am.* 146, 1568–1579.