IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 November 2022, accepted 28 January 2023, date of publication 15 February 2023, date of current version 27 February 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3245804*

RESEARCH ARTICLE

# Modeling of an Automatic Vision Mixer With Human Characteristics for Multi-Camera Theater Recordings

**ECKHARD STOLL[1,2], STEPHAN BREIDE[2], STEVE GÖRING[1], AND ALEXANDER RAAKE[1], (Member, IEEE)**

[1]Audiovisual Technology Group, Technische Universität Ilmenau, 98639 Ilmenau, Germany
[2]AudioVisual Media Center, South Westphalia University of Applied Science, 58644 Iserlohn, Germany

Corresponding author: Eckhard Stoll (eckhard.stoll@tu-ilmenau.de)

This work was supported by the Open Access Publication Fund of the Technische Universität Ilmenau.

**ABSTRACT** A production process using high-resolution cameras can be used for multi-camera recordings of theater performances or other stage performances. One approach to automate the generation of suitable image cuts could be to focus on speaker changes so that the person who is speaking is shown in the generated cut. However, these image cuts can appear static and robotic if they are set too precisely. Therefore, the characteristics and habits of professional vision mixers (persons who operate the vision mixing desk) during the editing process are investigated in more detail in order to incorporate them into an automation process. The characteristic features of five different vision mixers are examined, which were used under almost identical recording conditions for theatrical cuts in TV productions. The cuts are examined with regard to their temporal position in relation to pauses in speech, which take place during speaker changes on stage. It is shown that different professional vision mixers set the cuts individually differently before, in or after the pauses in speech. Measured are differences on average up to 0.3 seconds. From the analysis of the image cuts, an approach for a model is developed in which the individual characteristics of a vision mixer can be set. With the help of this novel model, a more human appearance can be given to otherwise exact and robotic cuts, when automating image cuts.

**INDEX TERMS** Automatic vision mixer, human characteristics, multi-camera theatre recordings.

*PRELIMINARY NOTE*

The term vision mixer is usually referred to both the device used to edit video and the person operating the device. The ambiguity is avoided in this paper by using the following terms:

- Vision mixing desk = the device used to edit video.
- Vision mixer (also abbreviated as VM) = the person who operates the vision mixing desk.

## I. INTRODUCTION

Professional recordings of live events are made with multiple cameras by qualified personnel. Camera work, picture direction and montage of theatrical recordings are artistic crafts that require training, experience, and skill. Therefore, it is difficult for amateurs to achieve acceptable results here. Without training in theory and practice, amateurs often do not know the design rules for aesthetically pleasing images and usually cannot follow movements of the performers quickly and competently.

This can be remedied by a production process in which medium shots and close-ups are obtained subsequently in post-production instead of during shooting. The recording would be done with high-resolution cameras (4K, 6K, 8K, or even more), which are fixed or only slightly panned and zoomed. Only long shots or medium long shots are recorded, which capture the entire action. This can also easily be done by inexperienced cameramen or camerawomen.

The associate editor coordinating the review of this manuscript and approving it for publication was Yue Zhang.

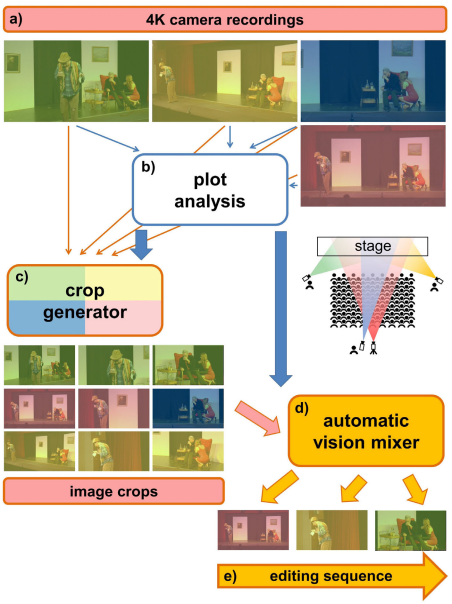


**FIGURE 1.** Schematic structure of an automatic editing system.

For example, in Figure 1 the concept of such an automatic editing system is shown, illustrated with the recording of a theater performance with a) four 4K cameras positioned left and right of the stage (green and yellow) and behind the audience (blue and red). In b) the camera signals pass through a plot analysis (white) that detects plot elements and logs them in a plot script, such as appearances and departures of protagonists, number of persons, positions and movements, face recognition, recognition of speaking persons, relationships of persons to each other (e.g., dialogues), etc. In c) the action script is used to search for and generate image sections (green, yellow, blue, red) from the 4K recordings, and in d) an automatic vision mixer (orange) is used to generate an editing sequence e). In our previously conducted work in [32], an online study is conducted to evaluate the quality and preference for three versions of the same scene with differently cropped image sections. There, it is investigated which compilation of cropping sections is preferred by subjects. In that study, the cropping of the three versions has been done manually using an image editing program. As a continuation, this paper deals with the automation of image cropping and the modeling of an automatic vision mixer which, as despite in Figure 1 automatically generates the cropping sequence and is supposed to show human properties (“humanize”) for the image editing. It is shown how choosing different parameters in the model change the cutting behavior in an application.

The paper is structured as follows: After the introduction in Section I, Section II gives an overview of the state of the art. Section III deals with the analysis of professional theatrical recordings. It discusses which broadcast series are suitable for analysis and the work of professionally working vision mixers (persons who operate the vision mixing desk) is examined. An overview of 13 selected recordings from five different vision mixers is given and how they are segmented into individual shots by automatic scene recognition. How image cuts can be set during speaker changes and whether image cuts occur during speaker changes or because actor movements can be distinguished by a plot analysis using the software OpenPose. Audio analysis can be used to determine speaker changes and speech pauses. The analysis of five different vision mixers is performed in Section IV. The different editing behavior of the vision mixers is presented and a model for automation is designed in Section V. In Section VI an algorithm is developed from the model and Section VII explains the application with examples. Discussion and summary conclude the paper in Section VIII.

## II. RELATED WORK

Due to the increase of automation in several fields, we outline in the next Section approaches for video editing. Lubart [21] discusses categories of human-computer interaction and how computers can be involved in creative work. Fully automatic techniques for editing videos are being developed for various fields. In Gandhi and Ronfard the focus is on the automatic detection and naming of actors on a stage [12]. In this work, a method is developed to distinguish the external appearance of clothes, which is implemented based on color differentiation. Gandhi et al. [11] apply person recognition to theater recordings for automatized generation of image details. Only one camera is used and the different obtained shots are extracted but not edited together. The system is then further developed in [13] to improve the tracking of actors in motion. In [15], the method is applied to 4K footage of dancers and multiple shots are output simultaneously on a split screen. Improvements, such as the use of a two-stage method (detection of timestamps for image cuts and optimization of crops for pans and zooms), were made by Rachavarapu et al. in [29]. In this work, the eye movement of a viewer is captured with an eye tracker. The image is optimized in x-position and zoom. The y-position is not changed and this restricts the algorithm so that faces or bodies could be cut off by the image boundary. Chen et. al. [5] investigate the computational complexity of optimal rectangle search in attention-based automatic image cropping. Fully automated image cropping approach based on a new model is proposed. Algorithms with low computational complexity are developed. Li and Zhang [19] generate image cropping using Collaborative Deep Reinforcement Learning (CDRL) trained by eye-tracking. Cropping is used to enhance the quality of experience (QoE) of 4K videos when played back on small screen devices such as smartphones [16]. Here the regions in the image that are frequently viewed are cropped and displayed in full format [7], [17].

Escobar and Parikesit perform an analysis of a theater video recording of a puppet show [8] and their approach uses difference frames of each two consecutive frames, the intensity of movement of the puppets is measured and assigned to narrative scene segments. Leake et al. are concerned with

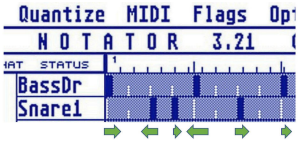


**FIGURE 2.** Notator SL humanize.

automated video editing for dialogues [18]. Using software such as OpenPose [3], [4], [31], [36] and OpenFace [14], face recognition and body tracking are applied and the video clips are matched to the textual dialogues in the script. Cuts are performed only when speakers change. Automatic segmentations of videos are performed for tutorial videos [34], for example, a semi-automatic video editing system will be developed to support the production of concise tutorials [6]. Automatic camera control of a single camera is used especially in amateur sports, because production with many cameras and camera crew is expensive. For example, Quiroga et. al. [28] film a basketball court with a fixed 4K camera and gain a lower resolution automated virtual camera to follow the game action. A similar method has been developed for ice hockey [26]. In comparison to basketball courts, typical soccer fields have a very large width. Therefore, different approaches are used to obtain a high-resolution 180° image as a basis [25], [27], [30] and specialized tracking algorithms track the game action [24], [33]. These methods are also used in other sports such as table tennis [22], tennis [9], or field hockey [23]. The methods and approaches listed have automated various subtasks in video production, but investigations of human characteristics in image editing and individual characteristics of professional vision mixers have not been incorporated, as they are used already for a longer time in audio. For example, for robot-like drums in the MIDI sequencer program Notator SL, the function ''humanize'' is developed as early as 1990 [2].

Figure 2 shows a section of the drum riff ''We will rock you'' quantized exactly to the bar units. The green arrows indicate how ''humanize'' shifts the beats to integrate human inaccuracies. The extent of the shifts is shown here enlarged rather than to scale to better illustrate the principle. The present paper is intended to contribute to the realization of a kind of ''humanize'' function for automatic image editing as well. This will be done using parameters obtained from professional theatre video recordings.

## III. ANALYSIS OF THEATRE VIDEO RECORDINGS

In this Section, professional theater recordings are selected and the editing behavior of vision mixers is analyzed. Speaker changes and speech pauses are determined by audio analysis and action analysis.

### A. SELECTION OF A TV THEATRE SERIES

To determine what constitutes human editing behavior, professional theatre video recordings are analyzed. A broadcast series of 13 recordings from the Ohnsorg Theater in Hamburg is used for the analysis because the processes are similar for each production. Thus, a larger number of recordings are available for the sample for data collection, which were recorded under similar conditions, such as stage size, number of cameras, camera positions, etc.

**FIGURE 3.** Camera arrangement in the Ohnsorg Theater. During rehearsals, the director (white) and vision mixer (orange) are in the front auditorium.

a) b)

**FIGURE 4.** Vision mixing desk a) during technical rehearsal and b) during recording.

At the Ohnsorg Theater in Hamburg, two fixed cameras are used at the front left (green, camera 1) and right (yellow, camera 2) of the stage, as well as two fixed cameras (blue, camera 3 and red, camera 4) in the audience, as Figure 3 shows. Such a setup is typically also used for recordings in non-professional environments. In order to even less disturb the audience, camera 3 and 4 are then usually placed behind the audience.

### B. PROFESSIONAL VISION MIXERS

The vision mixer is the person who operates the vision mixing desk and switches from one camera to another during production [35]. In Figure 3 the vision mixer (orange) can be seen next to the director (white) during rehearsals.

Figure 4 a) shows the vision mixing desk in front of the stage. Next to the vision mixer, the director observes the sequences and makes notes for corrections. For the recording, which is played with an audience, the vision mixer and director are in the OB truck (outside broadcast truck) Figure 4 b) standing outside the theater.

The basis for the recording is the script that the director has prepared in advance. There, the respective camera and shot size are determined for the individual text passages. The text in the script is indented to the extent that camera instructions can be handwritten in front of the text. Each change of speaker is separated from the previous speaker by a blank line. The vision mixer and the four camera operators have their own

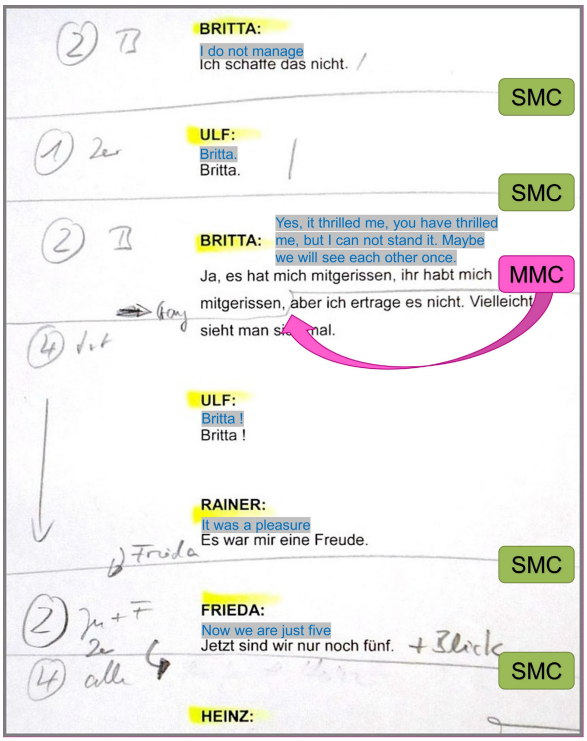


**FIGURE 5.** Script excerpt from the play "Dream dancers" with pencil instructions from the director.

scripts and individually write down their instructions in pencil so that they can be changed at any time.

Figure 5 shows a section of the vision mixer's script with original pencil marks of the director. Camera switches are marked by horizontal lines, the camera numbers to which the switch is made are circled. Who is to be seen in the picture is noted with letters (B = Britta) or abbreviations (alle = all, tot = German "Totale" = long shot, 2er = two shot). Furthermore, it can be seen that numerous cuts take place when the speaker changes, so that the person speaking in each case can be seen. For these cuts the term SMC (Speaker-change Motivated Cut) is introduced. If the speaker changes and the image cut occurs when the speaker changes, this cut is labeled as SMC. The term MMC (Movement-Motivated Cut) refers to cuts that occur due to movements of people. For example, a person is seen in close-up and that person or another person moves from one place to another on the stage. Or a person is performing or walk off. To capture the action, it is common to re-cut to a long shot. In Figure 5, such a cut is seen in mid-sentence. Britta in close-up announces that she is leaving and walks off while still speaking her sentence. In the workflow at the vision mixing desk, SMCs differ from MMCs. With SMCs, the text can be followed in the script and heard when the text passage is over. Then the vision mixer can decide where to cut in relation to the last word and the pause in speech. Here, editing could even be done with eyes closed. With MMCs, on the other hand, it is necessary to observe exactly when movements of the people on stage take place in order to make the cut. There is an individual assessment

**TABLE 1.** Thirteen analyzed plays and their respective vision mixers. (English translation of the German titles).

| Year | Play | VM |
|---|---|---|
| 1973 | Around Cape Horn | LOP |
| 1983 | The laurel wreath | BRW |
| 1983 | The cheerful gas station | BRW |
| 1984 | Short of Cash | LOP |
| 1989 | A man is not a man | SIB |
| 1997 | Rummy for three | SIB |
| 2013 | Lies have young legs | SVB |
| 2016 | Gossip in the stairwell | WIJ |
| 2016 | Country bumpkins | SVB |
| 2018 | When the cat's away | SIB |
| 2018 | Dream dancers | SVB |
| 2019 | A better gentleman | SVB |
| 2019 | Low German for beginners | SIB |

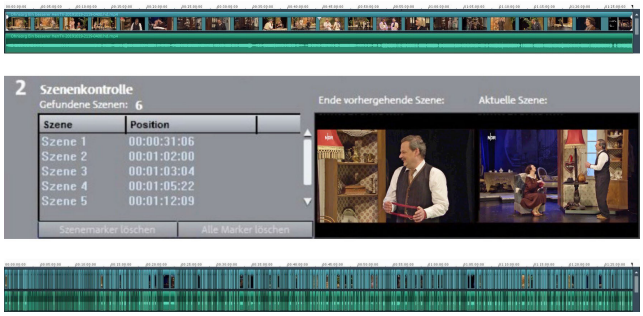


**FIGURE 6.** Automatic shot detection with Magix Pro. The recording (top) is analyzed (middle) and divided into individual shots (bottom).

of when, for example, a stand-up begins or a walk-off on stage happens. Often, the director also marks a text passage in the text where the cut to a long shot should take place because a new person appears on stage shortly afterwards or a person moves. Since this requires knowledge of the respective scripts, which are not available, only the analyses and results of SMCs are presented in this paper. Cuts due to speaker changes can be clearly identified and pauses in speech can be precisely timed by audio analysis.

### C. SELECTION OF TV RECORDINGS

For the analysis of SMCs, 13 TV recordings of the Ohnsorg Theater [10] are examined. In these, five different vision mixersare engaged.

Table 1 shows the play selection with the respective vision mixer (hereinafter abbreviated as VM), whose name is abbreviated with three letters. In the last 30 years, SVB and SIB were mainly engaged as vision mixers. A recording from 2016 was cut by WIJ and recordings from the 70s and 80s are analyzed that were cut by LOP and BRW.

To analyze the editing behavior for SMCs, the TV recordings are first segmented into the individual settings. The automatic scene detection function of the Magix Pro video software is used for this purpose. The "Scene Control" software function is applied to the complete video (compare Figure 6 top line). This detects shot changes (center right)


**FIGURE 7.** Breaking down a scene from "A better gentleman". The shots from the four cameras are assigned to four tracks. An image cut with speaker change is called SMC.

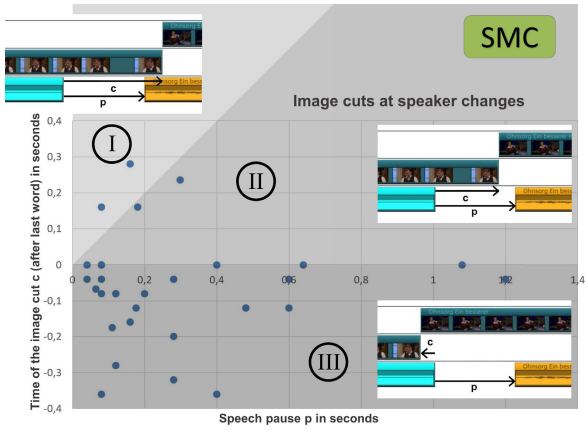


**FIGURE 8.** Image cuts of SMCs using the example of a scene from "A better gentleman". The horizontal axis shows the speech pause p. Vertically, the time of the frame c is plotted with respect to the last word of a person.

and enters the timecode in a list (center left). When the scene control for the complete video is finished, the settings are separated as individual objects in the timeline (bottom). The individual settings with the timecode of the cuts can be exported as an Edit Decision List (EDL), a list that is saved as text and contains timecodes of the In and Out points as well as information about source files, video and audio tracks, crops, etc.

### D. SMC VIDEO CUTS

For an initial analysis, Figure 7 shows the settings of a scene arranged on different tracks assigned to the four cameras. The long shots are placed on track 1, the semi-close-ups on which the woman and man can be seen together on track 2, and the close-ups of the man and woman on tracks 3 and 4. On the soundtrack below, the man's speech is shown in light blue, and the woman's speech is shown in orange, from the first word to the last. In this section of the scene, there are four SMCs and two MMCs.

To represent image cuts of SMCs, a "p-c diagram" is introduced, as Figure 8 shows. Each SMC is plotted as a point.

**FIGURE 9.** OpenPose body points.

Horizontally, the pause in speech $p$ (temporal distance) from the last word of one person to the first word of the other person is plotted. Vertically, the time of the image cut $c$ is plotted with respect to the last word of a person. The plot is divided into three different areas, which show how an image cut is set. For points in area I ($c > p$), the image cut occurs only after the first word of the other person. In area II ($0 < c < p$), the cut occurs during the pause in speech, and in area III ($c < 0$), a cut occurs before the last word.

### E. ACTION ANALYSIS WITH OPENPOSE

For the analysis of SMCs, MMCs must be automatically detected and sorted out. For this purpose, a pose recognition system is used. The OpenPose library was selected and is based on a neural network for human pose recognition [3]. It has been trained on about 25,000 images of over 40,000 people with annotated body joints [20] and shows good prediction performance.

Figure 9 shows an example of how OpenPose recognizes and displays body points of persons. OpenPose analyzes individual frames of video one after the other and outputs $x$, $y$ values for 25 body points of the recognized persons for each frame.

### F. DETECTION OF MMC

Using OpenPose, it is possible to detect and sort out MMCs so that the analysis can be applied to SMCs. Figure 10 shows

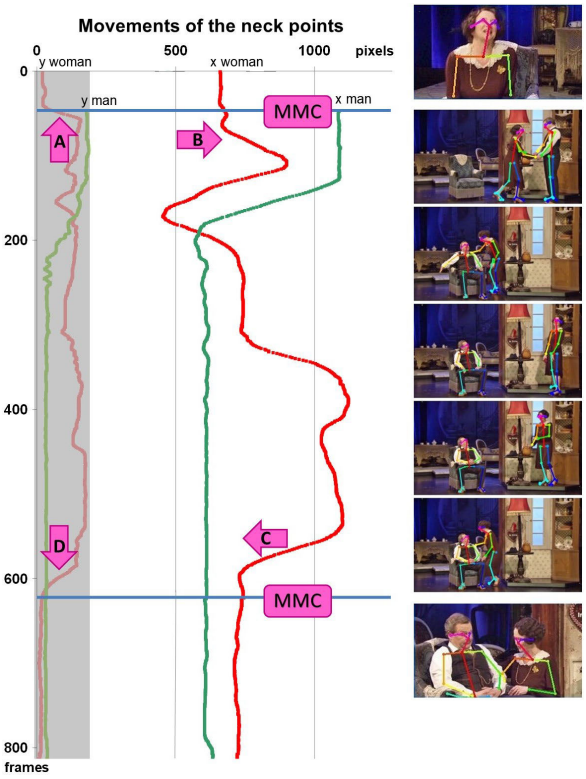


**FIGURE 10.** Movements on stage. x-values of the neck points of woman (red) and man (green) plotted over the time axis running vertically downwards.

a cut sequence of three shots. The positions of the neck points over time are shown in the diagram. The female x-values are red and the male x-values are green. In the gray field the corresponding y-values are shown. A rise of the woman is shown by the rapid increase of the y-values (arrow A), walking over to the man in the changes of the x-values (arrow B). Likewise, shortly before the second image intersection, the woman's x-values change by going over to the armchair (arrow C) and the y-values change when crouching down (arrow D). Image slices can thus be classified by examining changes in neck values just after or before an image slice. In case of strong changes, an MMC is present.

### G. SPEAKER ANALYSIS FOR SMC INVESTIGATION

When analyzing SMC, speaker changes and pauses in speech should be determined. This can be done with audio analysis using speaker diarization. Speaker diarization is the process of partitioning audio into homogeneous segments according to speaker identity. The open-source toolkit pyannote-audio [1] is used. pyannote-audio is an open source program library based on PyTorch focusing on audio analysis with machine learning and showed promising results in the evaluation.

An analysis of a scene in the play "A better gentleman" is shown in Figure 11. The colors red and blue are assigned to the two persons in the scene. For the analysis of speaker

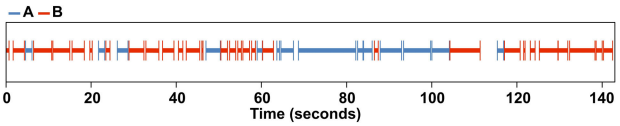


**FIGURE 11.** Speaker changes and speaking pauses of a scene in the play "A better gentleman", Ohnsorg Theater 2019.

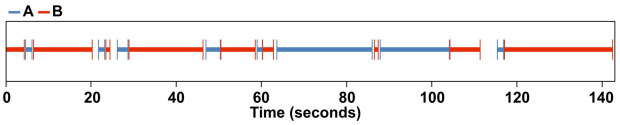


**FIGURE 12.** Summaries of the parts of speech.

changes, the pauses that can be heard within a person's speech are not relevant. The speech portions of the individual persons can therefore be summarized, as Figure 12 shows.

With a large number of actors in a play who have similar voices, persons can be mismatched. The following situations also lead to misinterpretations:

- Amazement: Ahh, Oh, Uhh
- Enthusiasm: Exclamations of jubilation, Wow, Hmmm.
- Cries of pain: Ow, Uhh
- Perplexity: Hm, Puh
- Hesitation sounds: Mhm, hmm
- Imitating or mimicking others
- Soundless utterances: Whispering, clapping hands, tapping shoulders
- Giggling, laughing, whistling
- Incidental sounds
- Audience noises
- etc.

Therefore, the results of the 13 examined plays are subjected to a manual check. It is checked whether in each case an SMC is actually present and whether the speaking pause was correctly determined. Inserts are also filtered out in which image cuts occur within a speech passage that cannot be assigned to a speaker change.

Figure 13 shows two insert cuts within a speech passage in the script that the director prepared in advance. Finally, 4446 manually checked SMC datasets are annotated from the total 9374 image cuts.

## IV. ANALYSIS OF DIFFERENT VISION MIXERS

It is investigated on the basis of the available 13 theater recordings whether a vision mixer shows a generally common editing behavior or whether different vision mixers show a different personal editing behavior. The editing behavior in the productions is analyzed for all 4446 SMCs and the respective timing of the image cutting related to the last word in each case is determined, as described in Section III-D. The cutting characteristics at SMCs of the five vision mixers LOP, BRW, WIJ, SIB and SVB are shown in Figure 14 in each case in the same representation as Figure 8. A color is assigned to each vision mixer (abbreviated as VM). At the bottom right,

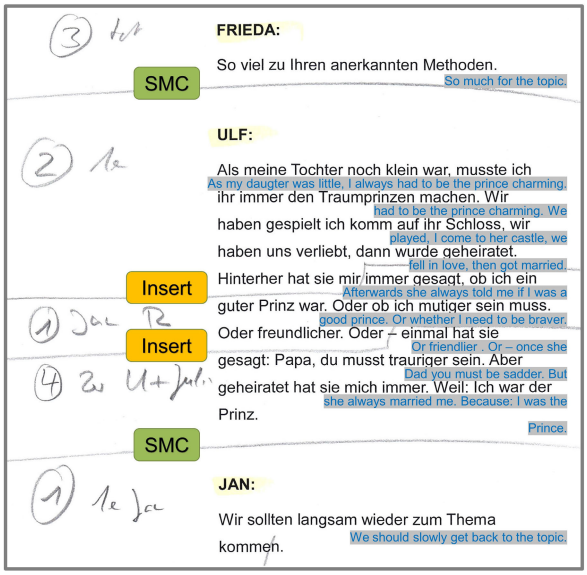


**FIGURE 13.** Script excerpt with image cuts without speaker changes (inserts).

the percentage distribution of SMCs for areas I, II and III is shown in the respective diagrams. The figures show that the cutting behavior of a given VM is very similar in the different productions. However, each VM seems to have a personal ''signature''.

In the case of VM LOP (green), the distribution among the areas does not show a clear preference in the two productions. Most of VM BRW's (yellow) image cuts, on the other hand, are in area I in both productions and thus behind the pause in speech. Preferred cuts in area III are found comparatively rarely. For VM WIJ (red) there is only one production. In this one, there tend to be more re-cuts after the pause (area I) than before the pause (area III). From VM SIB (yellow) and VM SVB (blue) there are four productions each, which are very similar in their cutting behavior, for VM SVB over a period of 6 years, and for VM SIB even over 30 years. Most of VM SIB's image cuts are in area I and thus behind the speech pause, whereas VM SVB prefers early image cuts. Most of his image cuts are in area III.

The different cut characteristics of the various VMs provide the viewer with different focal points in action capture. In cuts after the pause in speech, as preferred by VM BRW and VM SIB, the image of the previously speaking person is still visible when the other person has already started speaking. This allows the viewer to still see the final facial expression, e.g., a questioning eye-roll. The other person's direct reaction and introductory facial expressions, e.g., an indignant puffing of the cheeks, are not shown to the viewer. In VM SVB, it is the other way around. Because of the advanced cuts, the viewer can more frequently see the other person's direct reaction and introductory facial expressions even as the first person is finishing his or her speech. The final facial expression of the first person is not visible. Figure 14 also

**TABLE 2.** Analysis of the studied plays with number of SMCs and slopes of the respective straight lines m and y-axis intercepts b of the regression lines.

| Year | Play | vm | Min. | Shots | Shots/ Min. | SMCs | SMCs/ Shots | m | b |
|---|---|---|---|---|---|---|---|---|---|
| 1973 | Around Cape Horn | LOP | 106 | 456 | 4,30 | 181 | 39,69% | 0,0781 | 0,2235 |
| 1983(1) | The Laurel Wreath | BRW | 92 | 520 | 5,65 | 337 | 64,81% | 0,0611 | 0,2368 |
| 1983(2) | The cheerful Gas Station | BRW | 203 | 410 | 2,02 | 242 | 59,02% | 0,2171 | 0,3107 |
| 1984 | Short of Cash | LOP | 114 | 400 | 3,51 | 194 | 48,50% | 0,1239 | 0,0843 |
| 1989 | A Man is not a Man | SIB | 86 | 462 | 5,37 | 164 | 35,50% | 0,1257 | 0,2499 |
| 1997 | Rummy for three | SIB | 88 | 886 | 10,07 | 411 | 46,39% | 0,1749 | 0,1703 |
| 2013 | Lies have young Legs | SVB | 111 | 896 | 8,07 | 520 | 58,04% | 0,1482 | 0,0294 |
| 2016 | Gossip in the Stairwell | WIJ | 106 | 846 | 7,98 | 379 | 44,80% | 0,2152 | 0,1602 |
| 2016 | Country Bumpkins | SVB | 80 | 985 | 12,31 | 598 | 60,71% | 0,0743 | -0,0131 |
| 2018 | When the Cat's away | SIB | 93 | 924 | 9,94 | 344 | 37,23% | 0,1136 | 0,1786 |
| 2018 | Dream Dancers | SVB | 102 | 962 | 9,43 | 373 | 38,77% | 0,0554 | 0,0399 |
| 2019 | A better Gentleman | SVB | 89 | 814 | 9,15 | 430 | 52,83% | 0,0418 | -0,0106 |
| 2019 | Low German for Beginners | SIB | 88 | 813 | 9,24 | 273 | 33,58% | 0,1077 | 0,2544 |
| | | **Total** | 1358 | 9374 | 6,90 | 4446 | 47,43% | | |

shows the regression lines with the respective straight line Equation 1.

$$c = mp + b \tag{1}$$

In this equation, $m$ is the slope of the respective straight line and $b$ is the y-axis intercept.

Furthermore, Table 2 shows an overview of the determined values for $m$ and $b$. Also listed are the respective lengths of the recordings, number of shots, number of SMCs and their percentage relative to the total number of shots. Depending on the play, approximately between one-third and two-thirds of the settings could be determined to be SMCs. The number of settings and the percentage of SMCs vary because the plays differ in content. For example, in some plays there are more dialogue scenes where the actors do not move, whereas in other plays there is more movement.

A comparison of the regression lines is shown in Figure 15 a). Here, the different cutting characteristics of the vision mixers are shown in direct comparison. It can be seen that for each vision mixer the cutting characteristic is in an individual, color-coded zone range, with SVB (blue) cutting fastest, LOP (green) and WIJ (red) occupying a middle position and SIB (orange) and BRW (yellow) cutting slowest. It is noticeable that for all vision mixers towards larger speech pauses the cuts are set slightly later on average than for smaller speech pauses. All regression lines have positive slopes $m$, even if they are small with about 0.05 to 0.2. For example, for a slope of 0.1, if there is a pause in speech of one second, the cuts are on average one tenth of a second slower than if there is a direct change of speaker without a pause in speech. This behavior can perhaps be explained by the fact that in play scenes, where the speaking pauses tend to be longer during speaker changes, the vision mixers register this or that this is known to them through rehearsals and they thus know that they can take a little more time for the cuts. However, this is of little effect. Because if the vision mixers were to clearly adjust the image cut to the pause in speech, e.g., tend to place it in the middle of the pause in speech, then $m$ of about 0.5 would be detected, i.e., halfway between $m = 0$ and $m = 1$. This is because cuts exactly

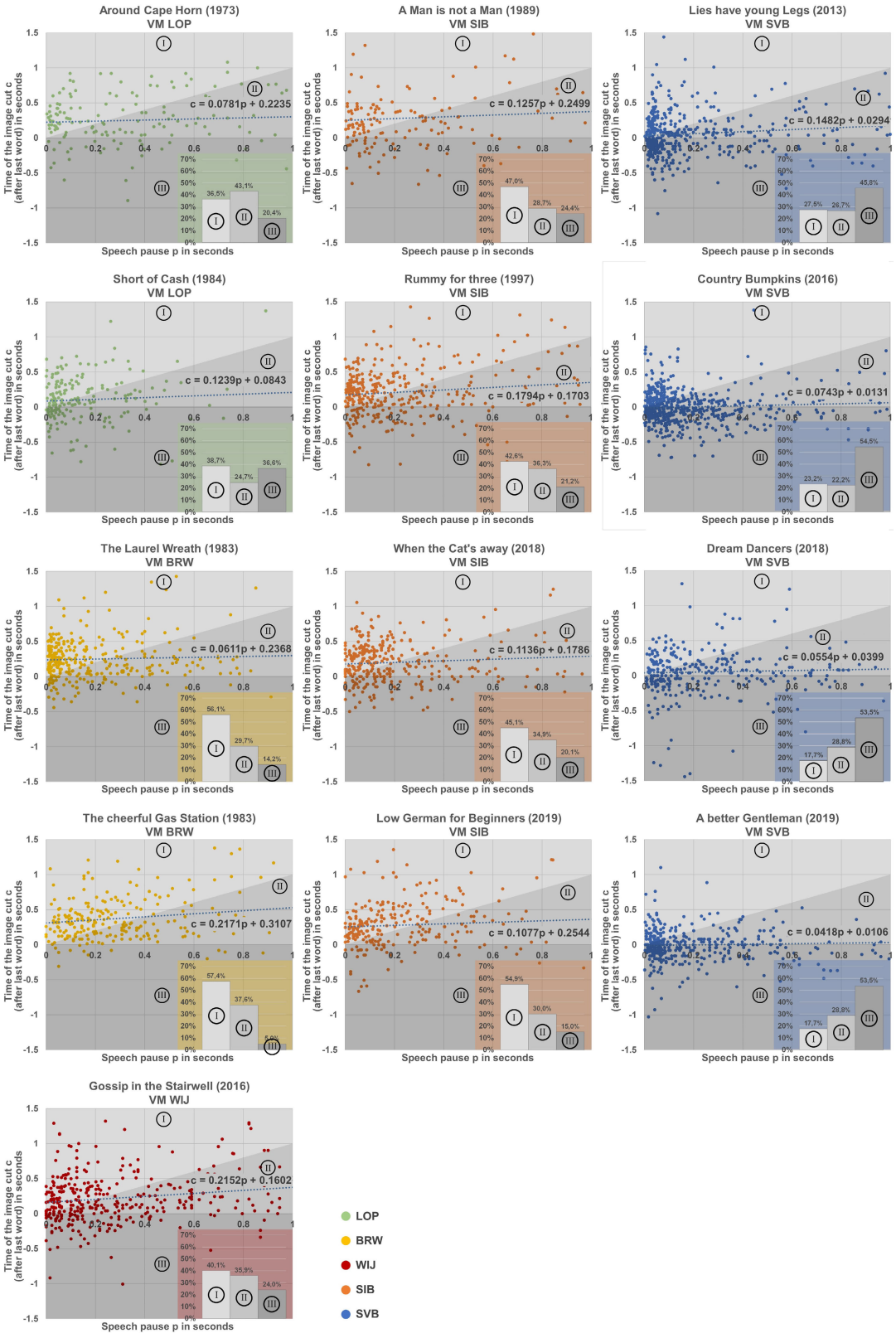


**FIGURE 14.** Editing behavior of the five vision mixers (five different colors), SMCs shown in p-c diagrams (see Figure 8). The horizontal axis shows the pause in speech p. Vertically, the time of the frame c is plotted with respect to the last word of a person. Area I: cuts after the speech pause, Area II: cuts during the speech pause, Area III: cuts before the last word.

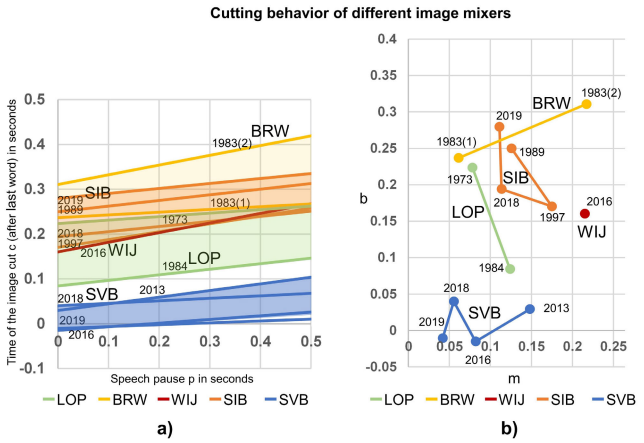


**FIGURE 15.** a) Regression lines of all 13 diagrams from Figure 14 and b) representation as m-b diagram.

at the beginning of the pause in speech lie on the straight line between zone II and zone III with slope $m = 0$ (cf. Figure 8), and cuts exactly to the end of the pause in speech lie on the straight line between zone I and zone II with straight line slope $m = 1$. For a better representation of the zone ranges, the values $m$ and $b$ are plotted in an m-b diagram in Figure 15 b). Each regression line is thus represented as only a single point. The points are connected chronologically for each vision mixer.

The differences in individual speed are shown here in the vertical direction. The y-axis intercept b represents the time of the image cut after the last word $c$, in the case of direct change of speaker without a pause in speech. Because if $p = 0$, then $c = b$ in equation (1). For VM SIB and VM SVB it can be seen in Figure 15 b) that the years of productions have no visible effect on the speed. Thus, for example, there is no tendency for newer plays to be cut faster than older ones. Rather, the b-values fluctuate from year to year, with no trend in any particular direction.

## V. VM-MODELING

The next step is to develop a model that reproduces the cutting behavior of a human vision mixer so that it can be used in an automatic montage (compare Figure 1). The user should be able to set, on a scale, whether the editing behavior reflects a rather pre-advanced editing as for vm SVB, or trailing edits as for BRW. For this purpose, the samples are analyzed and the descriptive parameters are determined such as mean value, variance but also skewness and excess, since the distributions are not normally distributed, as tests show.

### A. REGRESSION LINES

In the first step, the data sets of all productions of an vision mixer are cumulated from the 13 data sets into one common sample each, so that each SMC of an image mixer is weighted equally.

Table 3 shows the values of the cumulative samples of the five vision mixers. Figure 16 a) shows the cumulative

**TABLE 3.** Analysis of the cumulated data sets.

| vm | Min. | Shots | Shots/ Min. | SMCs | SMCs/ Shots | m | b |
|---|---|---|---|---|---|---|---|
| LOP | 220 | 856 | 3.89 | 375 | 43.81% | 0.1363 | 0.1350 |
| BRW | 295 | 930 | 3.15 | 579 | 62.26% | 0.1809 | 0.2562 |
| WIJ | 106 | 846 | 7.98 | 379 | 44.80% | 0.2152 | 0.1602 |
| SIB | 355 | 3085 | 8.69 | 1192 | 38.64% | 0.1378 | 0.2039 |
| SVB | 382 | 3657 | 9.57 | 1921 | 52.53% | 0.0997 | 0.0041 |
| **All** | 1358 | 9374 | 6.90 | 4446 | 47.43% | | |

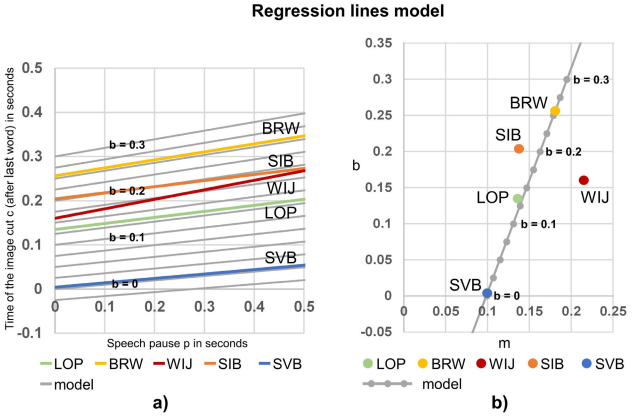


**FIGURE 16.** a) Regression lines model and b) representation as m-b diagram.

regression lines of the five vision mixers and in gray the straight line array of a model, depending on the selected y-axis intercept $b$.

The model is calculated in such a way that the lowest regression line of VM SVB and the highest regression line of VM BRW are exactly represented by the model and the intermediate straight line slopes are linearly interpolated as a function of $b$. In the m-b diagram in Figure 16 b), the model straight line thus runs through SVB and BRW. It can be seen that the cutting behavior values for LOP and SIB, which are close to the model straight line, are also well represented by the model. Only for WIJ the model does not fit so well. However, only one record of WIJ is available and the sample is small. The straight line chart of the model, where the y-axis intercept $b$ is selectable, is represented by the Equation 2.

$$c = m(b) \cdot p + b \tag{2}$$

The slope $m$ depends on $b$. BRW and SVB from Table 3 should be part of the straight line array and satisfy the Equation 2.

$$c = m_{BRW} \cdot p + b_{BRW} \tag{3}$$

$$c = m_{SVB} \cdot p + b_{SVB} \tag{4}$$

By transforming and inserting the values from Table 3, Equation 5 can be derived.

$$m(b) = \frac{b + 0.30558}{3.104558} \tag{5}$$

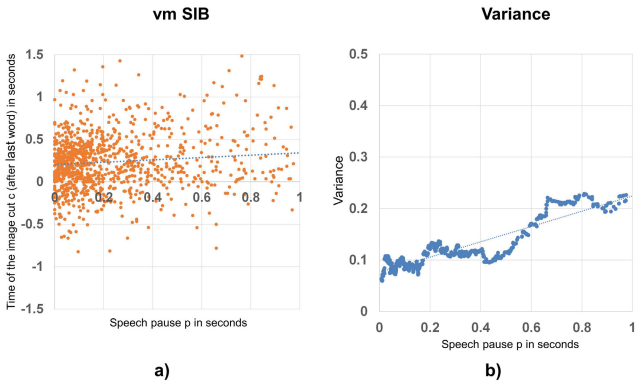


**FIGURE 17.** a) Cumulative sample SIB and b) p-dependent variance.

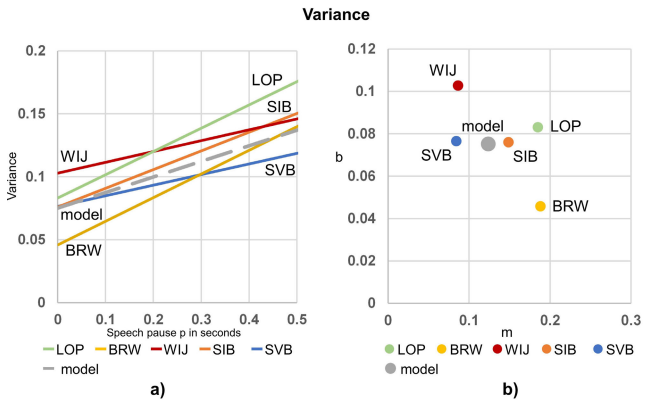


**FIGURE 18.** a) Variance model and b) m-b diagram.

Furthermore, the straight line array can be represented by Equation 6.

$$c = \frac{b + 0.30558}{3.104558} \cdot p + b \tag{6}$$

The user can now select any $b$, which should be approximately between 0 and 0.25 as shown in Table 3. With $b = 0$, a cut similar to SVB is the output, with 0.25 it is more similar to BRW.

### B. VARIANCE

The next step is to investigate which properties the individual cutting samples of the vision mixers (see Figure 14) have and how they can be simulated. A check with the Goldfeld-Quandt test shows that there is no homoskedasticity but heteroskedasticity. The variance of the interference terms is not constant, but dependent on the speech pause $p$.

Figure 17 a) shows the cumulative sample for SIB and b) the plot of the variance over the speech pause $p$. The local variance $s$ was determined by looking around a measured value at the respective interval from 49 measured values before to 50 measured values after the measured value and determining the local variance from these 100 values, see also Equation 7.

$$s_i = \frac{1}{100} \sum_{k=i-49}^{i+50} (x_k - \bar{x}_k)^2 \tag{7}$$

Figure 18 a) shows the trend lines of variance over p for the five vision mixers and b) the corresponding m-p diagram. For all of them it can be seen that the variance increases over the speech pause $p$. For the model, the y-axis intercepts and straight line slopes of the five straight lines are averaged, weighted by the number of SMCs of each vision mixer. SVB enters the model most heavily with 1921 SMCs, followed by SIB with 1192 SMCs (see Table 3). The resulting model straight line of variance $\sigma^2$, which depends on $p$, is plotted dashed in gray and is represented by the straight line Equation 8.

$$\sigma^2(p) = 0.12360\, p + 0.07520 \tag{8}$$

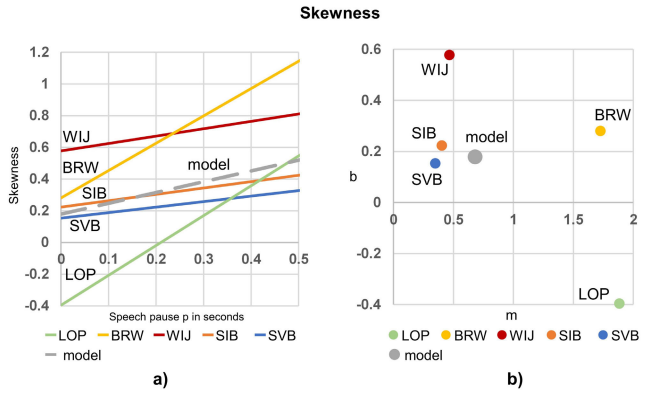


**FIGURE 19.** a) Skewness model and b) m-b diagram.

From the model equation, it can be seen that for small speech pauses close to zero, the variance is about 0.075. This means that about 68 % of the values lie in a range of about 270 ms. For speech pauses of half a second this range has grown to about 370 ms and for speech pauses of one second this range is about 450 ms.

### C. SKEWNESS

Another parameter to be examined is the skewness of the samples. Histogram analyses of the samples of the five vision mixers show that there is no symmetry. The measure for the asymmetry is the skewness $v$, which is determined like the variance in intervals with 100 measured values (in each case 49 before to 50 after the measured value) over $p$, since it depends on $p$.

Figure 19 a) shows the trend lines of skewness $v$ as a function of $p$ for the five vision mixers and b) the corresponding m-p diagram. The skewness is very different for the 5 vision mixers. For all vision mixers, the skewness increases as $p$ increases. For LOP, it is negative at first and then becomes positive during pauses in speech starting at about 0.2 seconds. For all others, it is positive throughout. It should be noted that the sample for LOP is very small with 375 SMCs compared to over 1000 for SIB and SVB. Furthermore, when determining the skewness, third powers are calculated. For the variance, it is only second powers. Therefore, the results are more

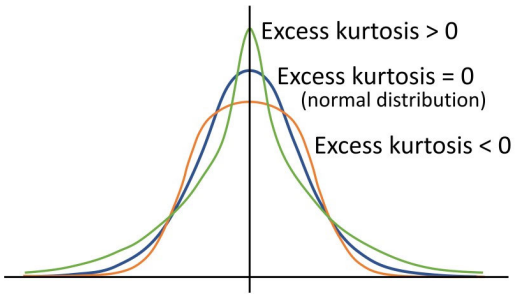


**FIGURE 20.** Excess kurtosis.

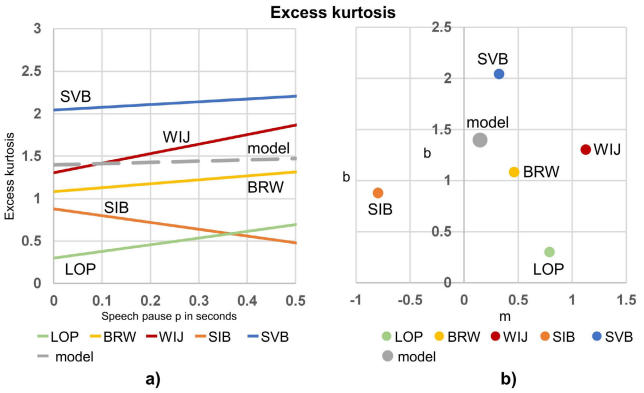


**FIGURE 21.** a) Excess kurtosis model and b) m-b diagram.

uncertain for small samples. But again, the model weights according to the number of available SMCs of the respective vision mixers. Thus, the similar measurements for SIB and SVB are strongly included in the model, as shown by the dashed model line in Figure 19 a).

For the modeling of the skewness $\nu$ the model straight line is then estimated by Equation 9.

$$\nu(p) = 0.68134\,p + 0.17874 \tag{9}$$

This estimated model line is drawn in gray in Figure 19 a) and in the corresponding m-p diagram Figure 19 b).

### D. EXCESS KURTOSIS

It is analyzed whether the samples of the vision mixers have the distribution of a normal distribution (excess kurtosis = 0), or whether they are more steeply curved (excess kurtosis positive) or more shallowly curved (excess kurtosis negative), as shown in Figure 20.

Figure 21 a) shows the trend lines of excess kurtosis as a function of $p$ for the five vision mixers and Figure 21 b) the corresponding m-p plot.

For all vm, the excess is positive, the distributions are more steeply curved than the normal distribution. The excess increases over the speech pause $p$, except for SIB. Analogous to variance and skewness, the model straight line of the excess $\gamma$ is determined and described by the straight line Equation 10.

$$\gamma(p) = 0.14825\,p + 1.39637 \tag{10}$$

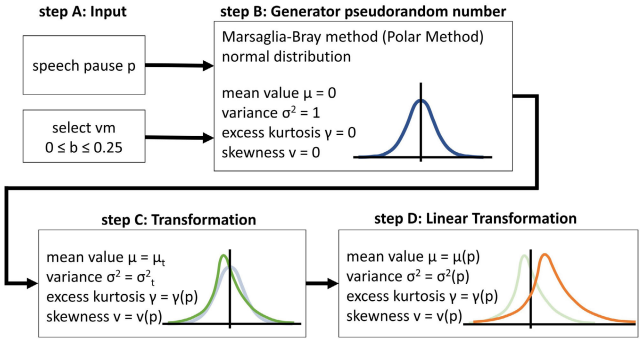


**FIGURE 22.** Algorithm steps.

## VI. ALGORITHM

From the estimated parameters of the model, an algorithm is designed that replicates the different editing characteristics of human vision mixers. The user can set a parameter $b$ in a range from 0 to 0.25. At $b = 0$, the editing behavior becomes similar to SVB with more advanced cuts in area III (see Figure 14), at $b = 0.25$ similar to BRW with most of the image cuts in area I and thus behind the speech pause. Then, for each SMC, a random variable is generated as a function of the speech pause and this is output as the time after image cut $c$ (see Figure 8). As described in Section V, the random variable has a p-dependent regression line, variance, skewness and excess kurtosis.

The algorithm that generates this pseudo-random variable is constructed in several steps, as Figure 22 shows. The input (step A) is, on the one hand the pause $p$ for which the algorithm is to be applied, and, on the other hand, the selected vm-parameter $b$. Using the Marsaglia-Bray method, a pseudo-random number is generated from a normally distributed random variable (step B). Afterwards, the excess and skewness are adjusted by transformations so that they correspond to the values of the model from Section V (step C). Corresponding transformations can be found at [37]. In step D, the desired mean and variance are adjusted by linear shift so that all parameters for the input variables $b$ and $p$ satisfy the Equations 6, 8, 9 and 10.

There is no need to pay attention to a special computing power, architecture or memory size because the computational complexity of the method is very small. The calculation of 100000 values according to the Marsaglia-Bray method in Python on standard PCs is usually specified with less than 0.2 seconds. However, less than 600 values are required per play according to Table 2. Step C and D are simple transformations and are also performed in milliseconds. Since the process is used in post-production and not in live operation, the time required is not critical.

## VII. MODEL APPLICATION

To illustrate the applications of the algorithms in Figure 23 an example is shown. A sequence from "When the cat's away" can be seen as it was cut a) in the original by vision mixer


**FIGURE 23. Algorithm in application. Selection of different image mixer "handwritings" by setting the parameter b in c) and d).**

SIB during live recording. Version b) shows how an automatic vision mixer without application of the vm-model exactly sets the cuts at the respective speaker changes and thus the editing behavior can appear robotic in the long run. If the VM model is applied, this is done in the editing program via an intervention into the Edit Decision List (see Section III-C). There, the beginning and end of each shot is entered chronologically in a table. These timecodes can be changed in the EDL according to the algorithm developed in Section VI. If $b = 0.25$ is set, which roughly corresponds to SIB (see Figure 16), then a random cut sequence can result as in c). In this case, this shows similar behavior to the original cut. If $b = 0$ is selected, then a different "handwriting" results as in d), which shows more advanced cuts and is similar to the cutting behavior of VM SVB. This cutting characteristic can be seen, for example, in Figure 7.

## VIII. SUMMARY

This paper presents a model for the cutting characteristics of SMCs for an automatic vision mixer. The personal preferences and typical characteristics of professional vision mixers in a live environment are analyzed and represented in a model. Five different vision mixers are examined, which show different characteristics in editing behavior. The basis are 13 productions of the Ohnsorg Theater Hamburg, which were recorded under similar production conditions. The analysis shows that numerous cuts take place during speaker changes so that the person speaking in each case can be seen. For these cuts the term SMC (Speaker change Motivated Cut) is introduced. The term MMC (Movement Motivated Cut) is introduced for cuts that occur due to movements of people. An examination of SMCs shows that a typical "signature" can be assigned to each vision mixer. For all vision mixers, regression lines, variance, skewness, and excess are determined as a function over the speech pause $p$, and a model is developed. The user can use a parameter $b$ to select whether, for example, image cuts should preferably take place before the speaker change or after the speech pause. An algorithm generates pseudorandom numbers according to this model and outputs time points of the image cut $c$ after the last word.

Improving the model with a larger data set and also investigating the editing behavior of more human vision mixers will be a future task. Furthermore, a larger number of productions considered per vision mixer would validate the model. However, collecting the data may prove difficult, since the credits of a recording usually list only the director and not the vision mixer by name. Production documents from older productions are often no longer archived at television stations. Therefore, an automated evaluation of a large number of productions without naming would only produce a model for an average vision mixer, which lacks the artistically important, individual characteristics. Since vision mixing is a creative, individual process, it is desirable to assign an individual profile to an automatic vision mixer, as is possible with the presented model.

## REFERENCES

[1] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.Audio: Neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7124–7128.

[2] *C-LAB team Notator User's Manual*, Atari ST Ser., C-LAB, Hamburg, Germany, 1990.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[5] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 507–515.

[6] P.-Y. Chi, J. Liu, J. Linder, M. Dontcheva, W. Li, and B. Hartmann, "DemoCut: Generating concise instructional videos for physical demonstrations," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2013, pp. 141–150.

[7] T. Deselaers, P. Dreuw, and H. Ney, "Pan, zoom, scan—Time-coherent, trained automatic video cropping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[8] M. E. Varela and G. O. F. Parikesit, "A quantitative close analysis of a theatre video recording," *Digit. Scholarship Humanities*, vol. 32, no. 2, pp. 276–283, 2017.

[9] M.-Y. Fang, C. K. Chang, N. C. Yang, and C. M. Kuo, "Robust player tracking for broadcast tennis videos with adaptive Kalman filtering," *J. Inf. Hiding Multimedia Signal Process.*, vol. 5, no. 2, pp. 242–262, 2014.

[10] Fernsehserien.de. *Ohnsorg Theater: Episodenguide*. Imfernsehen GmbH & Co. KG. 1998–2021. [Online]. Available: https://www.fernsehserien.de/ohnsorgtheater/episodenguide

[11] V. Gandhi, "Automatic rush generation with application to theatre performances," Ph.D. thesis, Math.-Inform., Grenoble Univ., Grenoble, France, 2014.

[12] V. Gandhi and R. Ronfard, "Detecting and naming actors in movies using generative appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3706–3713.
[13] V. Gandhi, R. Ronfard, and M. Gleicher, "Multi-clip video editing from a single viewpoint," in *Proc. 11th Eur. Conf. Vis. Media Prod.*, Nov. 2014, pp. 1–10.
[14] GitHub. (2020). *OpenPose: Real-Time Multi-Person Keypoint Detection Library for Body, Face, Hands, and Foot Estimation*. [Online]. Available: https://github.com/CMUPerceptual-Computing-Lab/openpose
[15] M. Kumar, V. Gandhi, R. Ronfard, and M. Gleicher, "Zooming on all actors: Automatic focus+context split screen video generation," *Comput. Graph. Forum*, vol. 36. no. 2, pp. 455–465, 2017.
[16] R. Kumar, P. Assuncao, L. Ferreira, and A. Navarro, "Retargeting UHD 4K video for smartphones," in *Proc. IEEE 8th Int. Conf. Consum. Electron.*, Sep. 2018, pp. 1–5.
[17] R. Kumar, L. Ferreira, P. A. A. Assuncao, and A. Navarro, "Retargeting 4K video for mobile access using visual attention and temporal stabilization," in *Proc. 9th Int. Symp. Signal, Image, Video Commun. (ISIVC)*, Nov. 2018, pp. 48–53.
[18] M. Leake, A. Davis, A. Truong, and M. Agrawala, "Computational video editing for dialogue-driven scenes," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–130, 2017.
[19] Z. Li and X. Zhang, "Collaborative deep reinforcement learning for image cropping," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 254–259.
[20] I. Lifshitz, E. Fetaya, and S. Ullman, "Human pose estimation using deep consensus voting," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 246–260.
[21] T. Lubart, "How can computers be partners in the creative process: Classification and commentary on the special issue," *Int. J. Hum.-Comput. Stud.*, vol. 63, nos. 4–5, pp. 365–369, Oct. 2005.
[22] H. Morimitsu, I. Bloch, and R. M. Cesar Jr., "Exploring structure for long-term tracking of multiple objects in sports videos," *Comput. Vis. Image Understand.*, vol. 159, pp. 89–104, Jun. 2017.
[23] Z. Musa, M. Z. Salleh, R. A. Bakar, and J. Watada, "GbLN-PSO and model-based particle filter approach for tracking human movements in large view cases," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 8, pp. 1433–1446, Aug. 2016.
[24] P. Ong, T. K. Chong, K. M. Ong, and E. S. Low, "Tracking of moving athlete from video sequences using flower pollination algorithm," *Vis. Comput.*, vol. 38, pp. 939–962, Jan. 2021.
[25] F. Perazzi, A. Sorkine-Hornung, H. Zimmer, P. Kaufmann, O. Wang, S. Watson, and M. Gross, "Panoramic video from unstructured camera arrays," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 57–68, 2015.
[26] H. Pidaparthy and J. Elder, "Keep your eye on the puck: Automatic hockey videography," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1636–1644.
[27] V. Popovic, K. Seyid, O. Cogal, A. Akin, and Y. Leblebici, "State-of-the-art multi-camera systems," in *Design and Implementation of Real-Time Multi-Sensor Vision Systems*. Berlin, Germany: Springer, 2017, pp. 13–31.
[28] J. Quiroga, H. Carrillo, E. Maldonado, J. Ruiz, and L. M. Zapata, "As seen on TV: Automatic basketball video production using Gaussian-based actionness and game states recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 894–895.
[29] K. K. Rachavarapu, M. Kumar, V. Gandhi, and R. Subramanian, "Watch to edit: Video retargeting using gaze," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 205–215, 2018.
[30] O. Schreer, I. Feldmann, C. Weissig, P. Kauff, and R. Schafer, "Ultrahigh-resolution panoramic imaging for format-agnostic video production," *Proc. IEEE*, vol. 101, no. 1, pp. 99–114, Jan. 2013.
[31] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1145–1153.
[32] E. Stoll, S. Breide, and A. Raake, "Towards analysing the interaction between quality and storytelling for event video recording," in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2020, pp. 1–4.
[33] M. Tiwari and R. Singhai, "A review of detection and tracking of object from image and video sequences," *Int. J. Comput. Intell. Res*, vol. 13, no. 5, pp. 745–765, 2017.
[34] A. Truong, P. Chi, D. Salesin, I. Essa, and M. Agrawala, "Automatic generation of two-level hierarchical tutorials from instructional makeup videos," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–16.
[35] P. Ward, *TV Technical Operations: An Introduction*. Boca Raton, FL, USA: CRC Press, 2013.
[36] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
[37] W. Von Zwet, "Convex transformations: A new approach to skewness and kurtosis," in *Selected Works of Willem Van Zwet*. Berlin, Germany: Springer, 2012, pp. 3–11.

**ECKHARD STOLL** received the degree in electrical engineering from the University of Karlsruhe. He carried out his first "Quasi" multi-camera productions in the 1980's. He is currently the Artistic Director of the Audio Visual Media Center, University of Applied Sciences Südwestfalen, teaches in the field of media production. He is also the Production Manager of Multi-Camera Productions. In cooperation with TU Ilmenau, he researches in camera based production technology. With one camera, four different performances of a play were recorded from four different camera positions, resulting in a multi-camera edit in post-production.

**STEPHAN BREIDE** received the degree in electrical engineering with a focus on communications engineering, communications systems engineering, and television engineering from the Technical University of Braunschweig. He teaches with the South Westphalia University of Applied Science in the field of communication services and applications. He is currently the Head of the Audio Visual Media Center. His focus is on multimedia applications and digital communication networks and the improvement of internet coverage in fixed wired broadband.

**STEVE GÖRING** received the B.Sc. degree, in 2012, the M.Sc. degree in computer science from TU Ilmenau, in 2013, and the Ph.D. degree in visual quality prediction using machine learning, in 2022. His focus is also on data analysis problems for video quality models and video streams. In 2016, he was with the Audiovisual Technology Group. He was worked with the Big Data Analytics Group, Bauhaus University Weimar. He is currently working as a Computer Scientist with the Audiovisual Technology Group, TU Ilmenau. His specializations are data analytics/machine learning, video quality, and distributed communication/information systems.

**ALEXANDER RAAKE** (Member, IEEE) received the Ph.D. (Dr.-Ing.) degree from the Electrical Engineering and Information Technology Faculty, Ruhr-Universitat Bochum, in 2005, with the book Speech Quality of VoIP. He has joined TU Ilmenau, in 2015, as a Full Professor, where he heads the Audiovisual Technology Group. Between 2005 and 2015, he held a senior researcher, an assistant professor, and a later associate professor positions with the An-Institut T-Laboratories, TU Berlin, a joint venture between Deutsche Telekom AG and TU Berlin, heading the Assessment of IP-Based Applications Group. From 2004 to 2005, he was a Postdoctoral Researcher with LIMSI-CNRS, Orsay, France. His research interests include audiovisual and multimedia technology, speech, audio and video signals, human audiovisual perception, and quality of experience. Since 1999, he has been involved with the ITU-T Study Group 12's standardization work on QoS and QoE assessment methods. He is also a member of the Acoustical Society of America, the AES, VDE/ITG, and DEGA.