Formant Distances and the Similarity Perception of Wind Instrument Timbres

Christoph Reuter[1], Isabella Czedik-Eysenberg[1], Saleh Siddiq[1], Michael Oehler[2]

[1]*Musicological Department, University of Vienna, Austria*

[2]*Musicological Department, University of Osnabrück, Germany*

[1]christoph.reuter@univie.ac.at, isabella.czedik-eysenberg@univie.ac.at, saleh.siddiq@univie.ac.at, [2]michael.oehler@uni-osnabrueck.de

Abstract

Do shorter distances between the formant positions of two sounds in a Euclidean space correlate with the impression of greater timbre similarity (and vice versa)? Is the numerical vector of formant positions a suitable basis for a precise computational classification of the involved instruments? Plotting the formants of 586 wind instrument sounds at all reachable pitches in ***ff*** and ***pp*** into a X/Y scatter diagram (X-axis: first formant, Y-axis: second formant), it is possible to visually distinguish musical instruments and their registers by their formant positions. In a listening test, 22 participants rated the (dis)similarity of 40 loudness-adjusted timbre combinations of wind instruments. Half of the stimuli contained sounds with extremely close/overlapping formant regions while the other half contained sounds with very distant formant regions. Moreover, with the help of machine learning methods, we tested the precision of the instruments' classification by means of their formants using 5-fold cross validation.

It turned out that timbres with very close formant positions are perceived as very similar while timbres with very distant formant positions are deemed very dissimilar. Based on these results we calculated a prediction model and compared it to a timbre similarity prediction model based on MFCCs.

When matching the formant positions of the examined timbres with their respective musical instruments we achieved a classification precision of 46.1% (cubic KNN). By adding a set of further timbre descriptors, we could increase the classification precision to 84.6% (quadratic SVM). The resulting confusion matrix corresponded very well with human mismatching of musical instrument timbres. In summary, the perception of timbre similarity in wind instruments was shown to be closely related to the sounds' formant positions. Regarding computational musical instrument classification, formant positions alone do not seem to be a sufficient criterion compared to other more powerful feature combinations, but regarding the evaluation of perceived timbre similarity, the distance between formant positions can serve as a helpful explanatory tool.

Background

It is a well-known fact in timbre research that the timbral character of a musical instrument as a whole cannot be sufficiently represented by the sound attributes of a single tone. Rather, musical instruments' timbres and the timbre of a single tone should be treated as two different concepts (Stumpf, 1926, p. 393; Siedenburg, Jones-Mollerup, & McAdams, 2016, p. 15). Nevertheless, in most publications about timbre research a single tone is treated as a prototypical representation of its source instrument. Only few concepts for a wider timbral description of an instrument through its entire dynamics and pitch range do exist. Beside Mel Frequency Cepstral Coefficients (MFCCs, e.g. Loughran, Walker, O'Neill, & O'Farrell, 2008) and Modulation Power Spectrum (MPS, e.g. Elliott, Hamilton, & Theunissen, 2013) formant areas were shown to be very useful for a more comprehensive description of musical instruments' timbres (since 1926, e.g. Stumpf, 1926; Schumann, 1929; Mertens, 1975; Meyer, 2009 etc.). Therefore, based on the formant areas of 586 wind instrument timbres in all reachable pitches and two dynamic levels we compiled a two-dimensional formant map (X-axis: formant 1, Y-axis: formant 2). The calculations were done using Praat (Boersma, & Weenink, 2013). In this map, the musical instruments (represented as point clouds in different colors) emerged to be clustered together by instrument, dynamics and register (see fig. 1). The calculated formant positions correspond to the descriptions found in the literature (e.g. Schumann, 1929; Mertens, 1975; Reuter, 1996; Meyer, 2009 etc.).

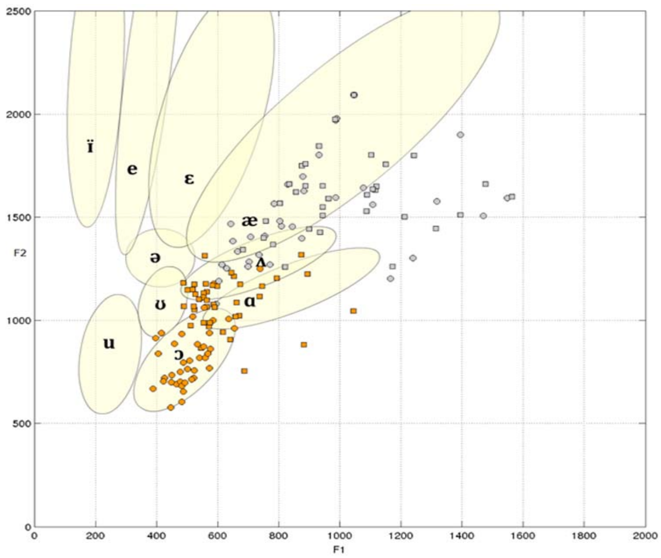

Figure 1. Formant map with the sounds of bassoon (orange) and oboe (grey) in all achievable pitches in *ff* and *pp* (Reuter, Czedik-Eysenberg, Siddiq, & Oehler, 2017).

Aims

These relatively well separated point clouds suggest that formant positions can perhaps serve as a useful and intuitive representation for a timbral description and classification of musical instrument.timbres. In addition, it indicates that sounds with short distances between their formants should be perceived as more similar than sounds with more distant formants. So, the aim of this study is to test the possible advantages and suitability of the formant paradigm for musical timbre similarity prediction as well as for timbre classification by investigating the following questions:

- Do shorter distances between the formant positions of two sounds correlate with the impression of greater timbre similarity (and vice versa)?

- Is the numerical vector of formant positions an suitable basis for a precise computational classification of the involved instruments? (Besides, which additional timbre features can improve the classification result?)

Computational description of perceived timbre similarity: Formants and MFCCs in comparison

In a listening test, 22 participants rated the (dis)similarity of 40 loudness-adjusted timbre combinations including wind instruments like flute, oboe, clarinet, bassoon, trumpet, trombone, French horn and tuba on a scale between 1 and 8 (8 = maximum dissimilarity). Half of the stimuli contained sounds with extremely close/overlapping formant regions while the other half contained sounds with very distant formant regions. Instrumental sounds were taken from the Vienna Symphonic Library (VSL) and adjusted to a matching ANSI-loudness level using the Genesis loudness toolbox in Matlab. The listeners' judgments were tested for correlations with the Euclidean distances between positions in the formant map. Indeed, a very significant correlation between the formant distances and the listeners' similarity judgments was found ($r = 0.759$, t-test with $p < 0.001$, confidence interval: [-3.1381, -1.8960]).

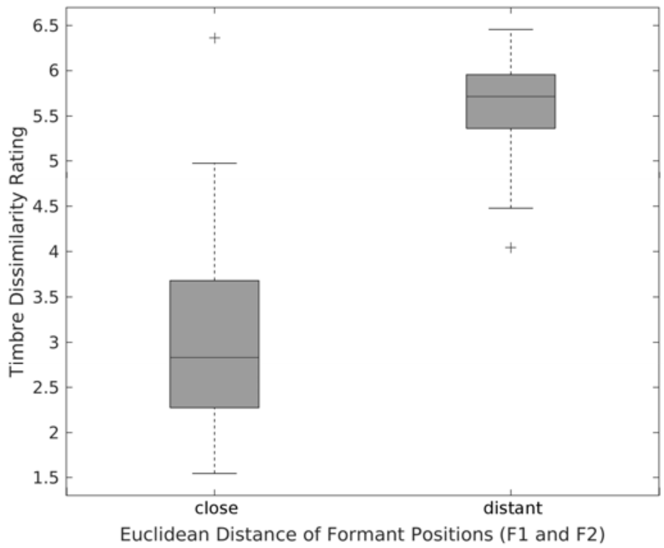

Figure 2. The distance of the formant positions (X-axis: close vs. distant) correlates strongly with ratings of perceived timbre similarity (Y-axis: 1 = very similar; 7 = very dissimilar).

Here, the first two formants (F1 and F2) have an almost equally strong linear relationship with the similarity scores and also have a very strong intercorrelation ($r = 0.9196$, $p < 0.0001$). In comparison, each of the first three MFCCs correlates weaker with the listeners' similarity scores, while the remaining MFCCs (4–13) show no significant correlations to the listeners' judgments at all. (Formants were calculated in Praat (Boersma & Weenink, 2013) and MFCCs in Matlab with the help of the MIRtoolbox (Lartillot & Toiviainen, 2007).)

Table 1. Correlation of individual formant positions and MFCCs with the perceived timbre similarity.

Timbre feature	r	p
F1	0.7514	< 0.0001
F2	0.7477	< 0.0001
F3	0.4227	< 0.0001
MFCC1	0.6384	< 0.0001
MFCC2	0.5959	< 0.0001
MFCC3	0.3513	< 0.05

A similar result can be seen when considering formant and MFCC combinations. In a two- or three-dimensional formant or MFCC space, the Euclidean distances based on the formants 1 and 2 correlate most strongly with the perceived timbre similarity ($r = 0.759$, $p < 0.0001$, see Fig. 3), while the Euclidean distances based on the MFCCs 1–3 give a slightly weaker but comparable picture ($r = 0.695$, $p < 0.0001$, see Fig. 4).

Table 2. Correlation of the Euclidean distances of formant (F1, F2, and F3) and MFCC combinations with perceived timbre similarity.

Timbre descriptors	r	p
F1 and F2	0.7591	< 0.0001
MFCCs 1, 2, and 3	0.6949	< 0.0001
MFCCs 1 and 2	0.6939	< 0.0001
F1, F2, and F3	0.6916	< 0.0001
MFCCs 1–13	0.6812	< 0.0001

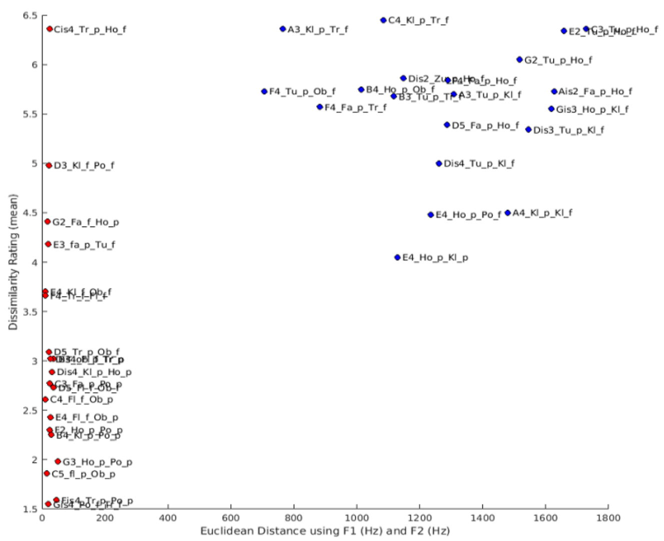

Figure 3. Scatter plot of the Euclidean distances of the first two formant positions (X-axis) and the perceived timbre similarity (Y-axis) (red: close formant positions, blue: distant formant positions).

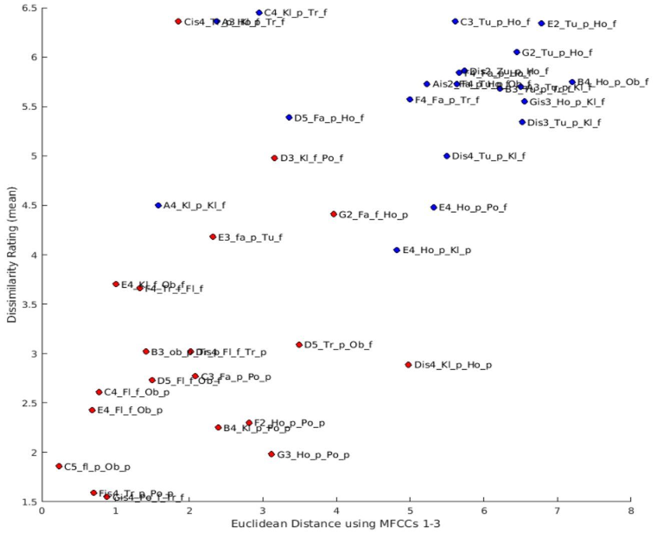

Figure 4. Scatter plot of the Euclidean distances of the first three MFCCs (X-axis) and the perceived timbre similarity (Y-axis) (red: close formant positions, blue: distant formant positions).

In a spider's web representation with the axes F1, F2, and F3 respective MFCC1, MFCC2, MFCC3 the triangles formed between the axes overlap very closely in case of a perceived timbre similarity (see Figure 5) while the distances of the triangle sides increase with increasing timbre similarity (see Figure 6).

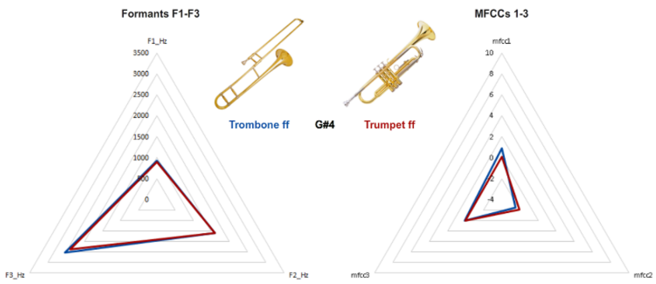

Figure 5. Spider's web representation with the axes F1, F2, F3 respective MFCC1, MFCC2, MFCC3 of the pair with the most perceived timbre similarity (trombone *ff* and trumpet *ff* on G#4).

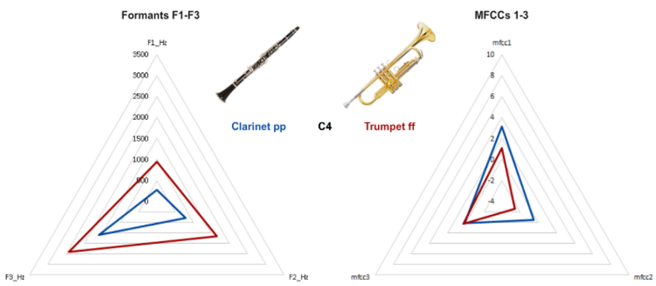

Figure 6. Spider's web representation with the axes F1, F2, F3 respective MFCC1, MFCC2, MFCC3 of the pair with the least perceived timbre similarity (clarinet *pp* and trumpet *ff* on C4).

A comparison of these results indicates that formants and MFCCs are similarly suitable for the determination of perceived timbre similarity. This is also suggested by the model of Darch et al. (2005), where formant areas can be derived on the basis of MFCCs.

Computational predictability of perceived timbre similarity: Formants and MFCCs in comparison

Regression models trained via machine learning (using 5-fold cross-validation) allow the prediction of perceived timbre similarities to a certain degree: Based on formants, the best result with a coefficient of determination (R^2) of 0.53 was found when using the absolute differences of F1 and F2 as features (see Figure 7).

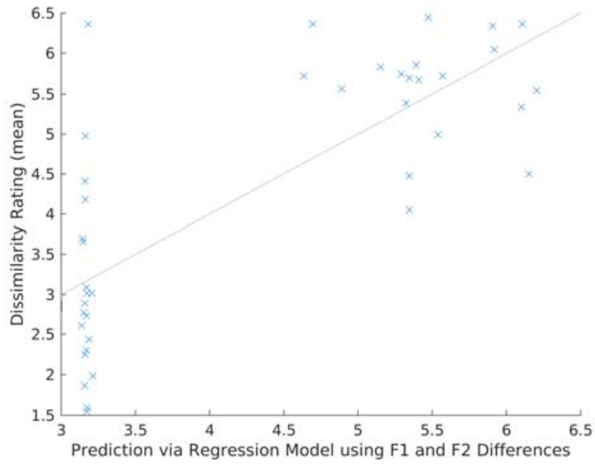

Figure 7. Prediction model based on F1 and F2 (R^2 = 0.53, RMSE =1.08, MSE =1.16, MAE =0.86).

Based on MFCCs a comparable result (R^2 = 0.56) can be obtained when usingMFCC1 and MFCC2 (see Figure 8).

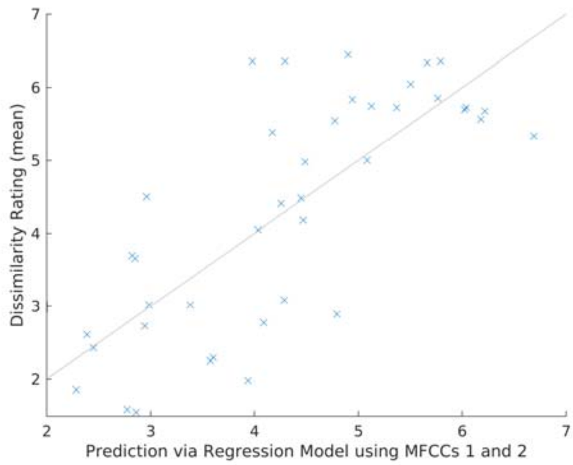

Figure 8. Prediction model based on MFCC1 and MFCC2 (R^2 = 0.56, RMSE= 1.05, MSE = 1.10, MAE = 0.81).

In other words, formants and MFCCs were approximately equally suitable for predicting timbre similarities. However, one has to take into account that other timbre features may also effectively predict perceived timbre similarities: For example, considering the distance of the spectral centroid of one sound to the other, it can be shown that with increasing distance between their values, the sounds' timbres are perceived as increasingly dissimilar (coefficient of determination (R^2) of 0.78, see Figure 9).

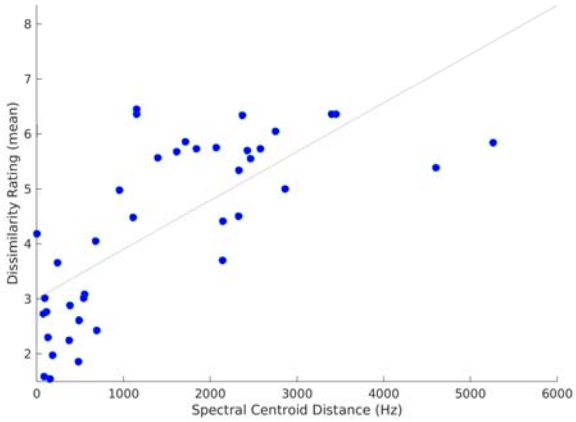

Figure 9. Prediction model based on spectral centroid (R^2 = 0.78, RMSE = 0.75, MSE = 0.56, MAE = 0.58).

Timbre classification based on formants

With the help of machine classification methods, timbre classes can be assigned to musical instruments with a precision of 46.1% based on the first three formants (F1 / F2 / F3, cubic k-nearest-neighbour classification) (see Figure 10).

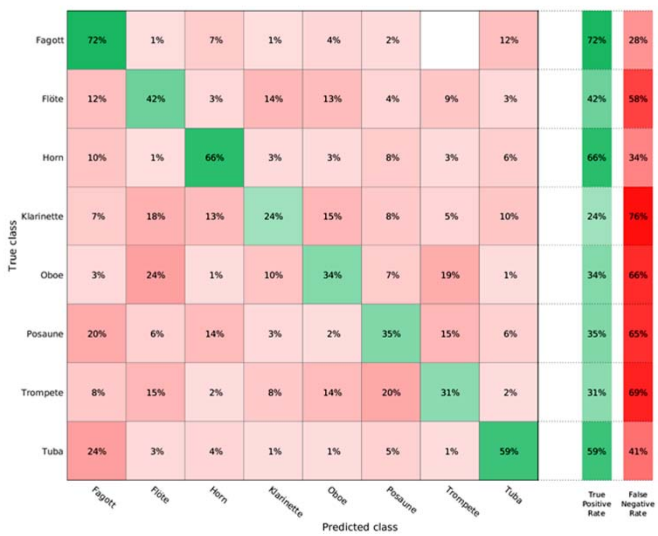

Figure 10. Confusion matrix based on F1, F2, and F3 (cubic KNN, classification precision of 46.1%).

When adding a set of complementary timbre descriptors to the formant positions, the precision of the instrument recognition can be increased to 84.6%. These additional timbre descriptors are attack time, spectral flux, roughness, brightness, spectral entropy, maximum RMS value, spectral centroid and unpleasantness. The last descriptor is implemented based on the results in Reuter, Oehler, & Mühlhans (2013), the rest is calculated with the help of Matlab MIRtoolbox (Lartillot & Toiviainen, 2007). The resulting confusion matrix shows plausible parallels to human perception, in the sense that instrument confusions frequently occurring in humans are also mirrored in the matrix. Examples are the confusions between flute, clarinet and oboe, between trumpet and trombone as well as in between tuba and bassoon timbres (see Figure 11).

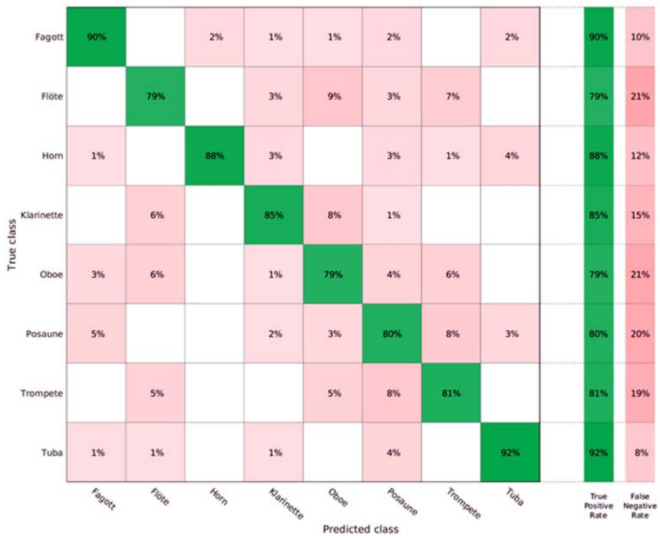

Figure 11. Confusion matrix based on F1, F2, and F3 as well as on attack time, spectral flux, roughness, brightness, spectral entropy, maximum RMS value, spectral centroid, and unpleasantness (quadratic SVM, classification precision of 84.6%).

Conclusion

The timbre similarity of wind instruments corresponds well to the proximity of the first two formants. A similarity prediction approach based on formant positions yielded results comparable to using MFCCs. Constructing classification models, we showed that a formant-driven instrument identification is possible, but does not lead to particularly high recognition rates. However, adding a small number of additional timbre descriptors yields an effective classification model that also shows correspondences to human perception. While formants, as sound descriptors for timbre similarity perception, strongly correlate with each other, individual MFCCs are more linearly independent of each other. However, the advantages of formants over MFCCs are: (1) formants need only two values to compactly describe a distinctive audible and therefore intuitively accessible spectral content and (2) they provide a solid foundation through more than 100 years of research history. For these reasons, formants can be considered as useful alternatives to MFCCs when analysing (wind) instrument timbres in music information retrieval.

Acknowledgements. This project is supported by funds of the Österreichische Nationalbank (OeNB, Anniversary Fund, project number: 16473). Special thanks to the Vienna Symphonic Library for supporting our work with two free "Vienna Super Package" licenses.

References

Boersma, P., & Weenink, D. (2013). *Praat: Doing Phonetics by Computer* [Computer program], Version 5.3.51, retrieved 2 June 2013 URL: http://www.praat.org/.

Darch, J., Milner, B., Shao, X., Vaseghi, S., & Yan, Q. (2005). Predicting Formant Frequencies from MFCC Vectors. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume I of V*. (pp. I/941–944). Pennsylvania, USA: Pennsylvania Convention Center/Marriott Hotel.

Elliott, T., Hamilton, L., & Theunissen, F. (2013). Acoustic Structure of the five Perceptual Dimensions of Timbre in Orchestral Instrument Tones. *Journal of the Acoustical Society of America*, 133(1), 389–404.

Lartillot, O., & Toiviainen, P. (2007). A Matlab Toolbox for Musical Feature Extraction from Audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx)*(pp. 237–244). France, Bordeaux.

Loughran, R., Walker, J., O'Neill, M., & O'Farrell, M. (2008). Musical Instrument Identification using Principal Component Analysis and Multi-Layered Perceptrons. In *Proceedings of the International Conference on Audio, Language and Image Processing ICALIP 2008* (pp. 643–648). China: Shanghai University.

Mertens, P.-H. (1975). *Die Schumannschen Klangfarbengesetze und ihre Bedeutung für die Übertragung von Sprache und Musik* [The Schumann Laws of Timbre and their Meaning for the Transmission of Voices and Music]. Frankfurt: Bochinsky.

Meyer, J. (2009). *Acoustics and the Performance of Music. Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instrument Makers*. New York: Springer.

Reuter, C. (1996). *Die auditive Diskrimination von Orchester-instrumenten* [Auditive Separability of Orchestral Instruments]. Frankfurt: Lang.

Reuter, C., Czedik-Eysenberg, I., Siddiq, S., & Oehler, M. (2017). Formanten als hilfreiche Timbre-Deskriptoren für die Darstellung von Blasinstrumentenklängen [Formants as helpful Timbre Descriptors for the Depiction of Wind Instruments Timbres]. In *Proceedings of the 43th. German Annual Conference on Acoustics „Fortschritte der Akustik“. DAGA 2017* (pp. 190–193). Germany: University of Kiel.

Reuter, C., Oehler, M., & Mühlhans, J. (2014). Physiological and acoustical correlates of unpleasant sounds. In *Proceedings of the Joint Conference ICMPC13/APSCOM5* (p. 9). South Korea: Yonsei University, Seoul.

Schumann. K.E. (1929). *Physik der Klangfarben* [Physics of Timbres]. (Habilitation Thesis). University of Berlin, Berlin, Germany.

Siedenburg, K.; Jones-Mollerup, K., & McAdams S. (2016). Acoustic and Categorical Dissimilarity of Musical Timbre: Evidence from Asymmetries between Acoustic and Chimeric Sounds. *Frontiers in Psychology*, 6:1977.

Stumpf, C. (1926). *Die Sprachlaute* [The Sounds of Voice]. Berlin: Springer.