Athanasios Lykartsis, Stefan Weinzierl, Volker Dellwo

Speaker Identification for Swiss German with Spectral and Rhythm Features

Lykartsis, Athanasios; Weinzierl, Stefan; Dellwo, Volker (2017): Speaker Identification for Swiss German with Spectral and Rhythm Features. In: 2017 AES International Conference on Semantic Audio
http://www.aes.org/e-lib/browse.cfm?elib=18753

Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

Speaker Identification for Swiss German with Spectral and Rhythm Features

Athanasios Lykartsis[1], Stefan Weinzierl[1], and Volker Dellwo[2]

[1]*Audio Communication Group, Technische Universität Berlin, Germany*
[2]*Phonetics Laboratory, Universität Zürich, Switzerland*

Correspondence should be addressed to Athanasios Lykartsis (athanasios.lykartsis@tu-berlin.de)

ABSTRACT

We present results of speech rhythm analysis for automatic speaker identification. We expand previous experiments using similar methods for language identification. Features describing the rhythmic properties of salient changes in signal components are extracted and used in an speaker identification task to determine to which extent they are descriptive of speaker variability. We also test the performance of state-of-the-art but simple-to-extract frame-based features. The paper focus is the evaluation on one corpus (swiss german, TEVOID) using support vector machines. Results suggest that the general spectral features can provide very good performance on this dataset, whereas the rhythm features are not as successful in the task, indicating either the lack of suitability for this task or the dataset specificity.

1 Introduction

The efficient description of speech rhythm is a challenging task which has been solved with limited success so far. The reason for this is the difficulty to define, measure and quantize what exactly constitutes speech rhythm. However, many studies up to now have shown that the rhythmic characteristics or even the general temporal evolution of speech, together with other factors, play an important role in the perception of language, especially for tasks such as speaker identification (SID) and language identification (LID), or even speech intelligibility [1, 2, 3, 4, 5, 6]. Therefore, further research on the subject could serve determining the important constituent elements of speech rhythm which contribute to language and speaker variability; and the creation of better features for speech processing.

Concerning speech rhythm feature extraction, the most influential studies have been performed in linguistics and phonetics. The basic assumption of those approaches is that rhythm-related speech phenomena take place on the level of the duration of intervals, phonemes, syllables, words and phrases. Therefore, metrics such as ΔC, $\%V$, *nPVI* and *VarcoC* [7, 1, 2, 8, 3] have been developed to capture the variability in the duration of syllables or consonant-vowel cluster intervals. However, recent observations [9, 10, 11] also criticize that those metrics are not necessarily characteristic of (solely) language variability. One novel approach for speech rhythm description are the attempts to describe speech rhythm related periodicities inherent in the signal. Such approaches for rhythm-based LID have been introduced based on automatic segmentation and feature extraction [12, 13, 14, 15, 16], low-frequency

periodicity analysis [17, 18, 6] and lately with methods borrowed from the field of Music Information Retrieval (MIR), e.g. with the beat histogram [19]. When looking specifically at the task of speaker ID, such approaches have been applied only to a lesser extent. However, recent studies [4, 20, 17] on speaker idiosyncratic speech rhythm features point toward the need to experiment with novel rhythm description methods. Standard SID approaches using machine learning methods with the help of basic features [21] and i-vectors [22, 23, 24, 25, 26] have provided good performance results in speaker recognition. Especially the i-vector approach in combination with Deep Learning has shown very high performance [27, 28, 29, 30]. These methods, however, are computationally complex and expensive and require a large amount of data for the building of the Universal Background Model (UBM), as well as for the training of the Deep Neural Nets (DNNs). Furthermore, it is largely unclear which features function well and why, as well as how they relate to specific qualities of speech (e.g. rhythm), with rhythm related features almost totally absent. Finally, the methods are applied to datasets which are not widely accessible since they are very expensive to obtain or only available in a challenge context (e.g. the NIST datasets), making the reproducibility of results difficult.

In this paper, we have therefore applied a novel method to extract speech rhythm related features for SID using the data of the swiss language TEVOID corpus [17] in order to determine if the proposed rhythmic features can be as successful for SID as they have been for LID [19]. Those features were selected, since speech rhythm metrics have been shown to provide interesting results for speaker identification. It is therefore interesting to evaluate our approach to rhythm features on the same dataset in order to check for consistencies or differences and draw conclusions about the features. At the same time we will test standard features in audio content analysis [31] as well as from speech processing - Shifted Delta Cepstral Coefficients (SDCs) and Mel Frequency Cepstral Coefficients (MFCCs) - as a baseline. We chose this dataset since it was accessible and it has been analyzed using the speech rhythm metrics [17], to which we wanted to compare our approach.

The paper is structured as follows: The feature extraction method is shortly described. The steps of the experimental setup feature evaluation for the TEVOID corpus are presented and discussed. Finally, conclusions and perspectives for further research are given.

2 Methods

2.1 Feature Extraction

For the extraction of rhythmic features for the SID task, we utilize the method proposed in [19], where five different novelty functions, i.e. temporal trajectories of different signal properties or their derivatives [32], are calculated and used as the basis for the creation of beat histograms, similar to the periodicity representations in [33, 34, 35]. We extract five such novelty functions:

- **Spectral Flux (SF),** following strong changes in (wideband) spectral properties.
- **Spectral Flatness (SFL),** indicating whether the signal is strongly tonal or noisy.
- **Spectral Centroid (SCD),** giving information about the spectral center of weight.
- **RMS Amplitude (RMS),** the standard amplitude/level information of the signal.
- **Fundamental Frequency (F0),** following the basic F0 information in the speech signal (extracted using the harmonic product sum method, see [31]).

The interested reader can refer to [31] for more information on the mathematical definition and the properties of those audio features. A beat histogram from the temporal trajectories of those features (in a given texture frame, i.e., a smaller window of the whole audio file) is extracted by computing the scaled autocorrelation function for frequencies from 0.5 to 10 Hz. From the beat histograms, the following statistical and other features (subfeatures) are extracted in turn (95 in total, resulting from 5 novelty functions and 19 subfeatures):

- **Distribution statistics:** Mean (ME), Standard Deviation (SD), Mean of the Derivative (MD), Standard Deviation of the Derivative (SDD), Skewness (SK), Kurtosis (KU), Entropy (EN), Beat Histogram Centroid (CD) and High Frequency Content (HFC).
- **Peak related:** Strength and Position of the First and Second Strongest Peak (P1, A1, P2, A2), Ratio (RA) of the Strength of the first Peak (A1) to that of the Second one (A2), Peak Centroid (P3), Sum (SU) and Sum of Beat Histogram Power (SP).

Almost the same parameterization as in [19] was used here; all files were resampled to 22050 Hz, and a window of 512 samples with an overlap of 75% was applied. A texture window of 4 seconds with a 50% overlap was used for creating several beat histograms, which were then averaged across the whole audio file. Other values for those parameters were considered, but those provided the best results. Apart from the rhythmic features, spectral ones were extracted by calculating the feature value over analysis frames of a Short-Time-Fourier-Transform (STFT) with the same parameterization as above for the whole audio file. We included the following 34 features (for more information, see [31]):

- **Spectral Shape and Change:** Spectral Flux (SF), Spectral Centroid (SCD), SDCs derived from the MFCCs ($1-13$, resulting in 13 SDCs in the $7-1-1-1$ setting, see [36] for more details), the MFCCs themselves, Spectral Flatness (SFL) and Spectral Spread (SSP).
- **Tonal:** Spectral Tonal Power Ratio (STPR) and Zero Crossing Rate (ZCR).
- **Envelope:** Root Mean Square Amplitude (RMS) and Envelope Max (EMX).

2.2 Classification

In order to perform supervised classification we have used the Support Vector Machines (SVM) [37] algorithm in a MATLAB implementation with a Radial Basis Function (RBF) kernel for a multi-class setting. The hyperparameters for the RBF kernel (C, γ) were determined through a grid search procedure. For all experiments, a 10-fold cross-validation took place. This means that the dataset was randomly separated in 10 equally large subsets (folds), out of which 9 were used for training and 1 for testing (validation). This procedure was repeated 10 times (corresponding to the number of the folds) and the average accuracy over all trials was computed. This represents a common way to perform machine learning experiments (e.g. in the MIR community) and assures that no skewed results are produced because of a single random advantageous or disadvantageous partitioning of the dataset. When the dataset is small, this could lead to problems with insufficient training material, which is why we chose a partitioning with relatively many folds (10). Z-score standardization was conducted prior to classification, separately for the training and test set. The *accuracy*, as the number of correct classifications for one class, to the number of overall classifications, was used to evaluate classification performance. We are primarily interested in this measure, as we are performing a 1-vs-1 multiclass supervised classification setting - that is, for each speaker pair, classifications are performed (in each fold), as we wish to know how well the algorithm can distinguish one speaker in comparison to another, and not to all others together (as in a 1-vs-all setting), since we can then interpret misclassifications in an easier way. The final result is calculated by summing the individual results for each class. This way we can also detect effects misclassified classes, which would point at speakers having similar properties (as measured through our features) or some speakers having not enough variance to stand out in comparison to any other class.

2.3 Datasets

For the speaker ID task, the TEVOID corpus was used [17]. It contains sixteen spontaneous utterances from sixteen male and female (50% for each category) Swiss German speakers (i.e. 256 utterances in total) transcribed and read by all speakers, resulting in 4096 sentences. The audio signal quality is high, and the corpus has already been analyzed [17] using many established speech rhythm metrics ($\%V$, $\Delta V(ln)$, $\Delta C(ln)$, $\Delta Peak(ln)$) and was found to contain considerable between-speaker variability, even when strong within-speaker variability was introduced. In this sense, it is expected that the speakers could be identified from a supervised learning algorithm successfully. It must be mentioned, however, that the database is relatively small, which could make the generation of reliable results difficult. Furthermore, the fact that the dataset deals with only variety of the german language (swiss german) could lead to results of the SID experiment might be specific for this language.

3 Results

Using the rhythm feature set (see the confusion matrix, Fig. 1), it was observed that all speakers are identified with an accuracy above chance level (Accuracy $>$ Prior $= 6.25\%$) while speakers S4, S7, S8, S10, S12 and S16 are identified with relatively low accuracy, below

20%. On the other side, two out of sixteen speakers are identified with relatively high absolute accuracy (S2 with 53.9% and S3 with 54.7%), three others with moderately good accuracy (S1, 36.3%, S9, 30.9% and S14, 31.6%) and for the rest of the speakers an accuracy of 20 to 30% is achieved. The average accuracy is 26.95%, which is more than four times greater than chance classification accuracy, but still in absolute values not entirely satisfactory. Using the spectral feature set (Fig. 2), the results where unambiguous: the overall performance was 87.6%, without much variation between speakers (around 10%). Speakers S3, S6, S10, S14, S15 are identified with an accuracy above 90%. This points towards the fact that simple, spectral features capture very speaker-specific characteristics such as voice timbre or F0. This confirms findings from other SID studies [22, 21, 23]. When combining both feature sets (Fig. 3), an 82.3% average accuracy is reached, which does not show much variation between speakers. This shows two interesting effects: First, accuracy actually decreases when using spectral features together with the rhythm related ones, hinting towards the fact that when using all the features with the same SVM classifier, the determination of a good class separation becomes more difficult. A similar effect was observed when using the linear SVM and the kNN classifiers, however with lower accuracy. Secondly, the variation pattern follows that of the spectral features, showing that they dominate in the task.

4 Discussion

The results presented in the previous section give a mixed picture. Using the rhythm features, it can be seen that the overall performance (as measured by accuracy) on the TEVOID corpus is relatively low (26.95%). This points towards the fact that the features do not necessarily capture speech rhythm in the same way as the rhythm metrics do, since when using the latter, it could be shown that between-speaker rhythmic variability in this corpus is robust and even with respect to certain kinds of within-speaker variability [17, 38]. However, the fact that recognition stays well above the prior in most cases is encouraging with respect to the features capturing some speaker related rhythmic variability. The spectral features have achieved a very high overall performance (87.6%) showing that SID with good results is possible even with the use of an uncomplicated, fast feature extraction scheme, opting for their use in future experiments and applications.

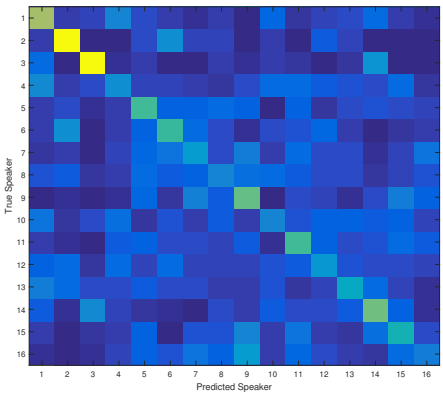

Fig. 1: *Confusion Matrix for the TEVOID corpus, rhythm features (16 speakers).*

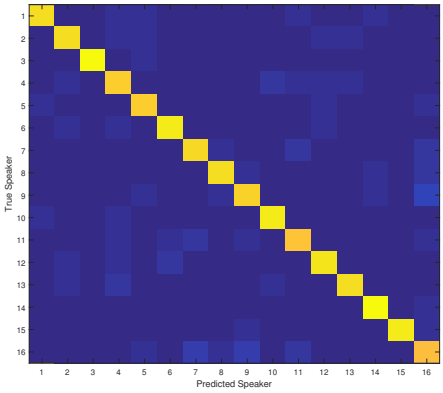

Fig. 2: *Confusion Matrix for the TEVOID corpus, spectral features (16 speakers).*

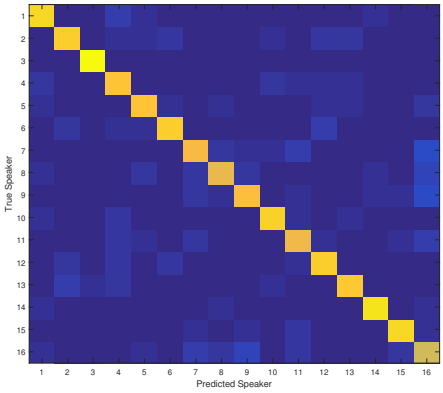

Fig. 3: *Confusion Matrix for the TEVOID corpus, combined features (16 speakers).*

The reasons for the unsatisfactory performance of the rhythm features might lie in the specific variety of swiss german in the corpus, which might be a special, difficult case to analyze in terms of rhythmic variability. Also interesting is the fact that specific users (two in particular) are identified with relatively high accuracy. This is a hint towards the assumption that our rhythm features capture very specific rhythmic patterns of certain individuals, which might have to do with the specific dialect of german, rate of speech or elicitation method (as the rhythm features did not perform well on spontaneous speech for LID either, see [19]) although those parameters have to be investigated further. A listening probe into the speaker characteristics of the best and worst cases did not reveal any rhythm-specific reasons for them performing better or worse, other than the fact that speakers S2 and S3 speak relatively slowly and somewhat more clearly. In this context, the investigation of perceptual similarities in speech rhythm between speakers through a listening experiment could also be helpful. To summarize, the fact that the results are significantly above the chance shows that rhythm features can indeed be helpful for SID but need to be further refined for use in such tasks. Possible problems could include the temporal resolution of the rhythm features (which could be adjusted to, e.g., fit the speaker rate) or the elicitation method. All of the above imply that SID (in contrast to LID) is much better served by just using spectral features, as they apparently capture a great part of speaker specificity. This might be a result of different speakers of the same language having very different voice timbre characteristics, which are readily captured through features such as the MFCCs, the SDCs and similar ones. In general, the high performance of the spectral features is similar to results shown elsewhere (e.g., the studies that use i-vector methods derived from MFCCs and SDCs [22, 23, 24, 25, 26, 27, 28, 29, 30]), which achieve error rates of 5% or lower on various datasets, ranking just a bit higher than our spectral features, but with a much more effortful analysis. On the other hand, speaker specific rhythm characteristics might either be absent (in general or for the dataset and language used here), very confounded with other sources of rhythm variability (such as elicitation method, emotional speech) or might just not be captured through our rhythm extraction method. Since using those rhythm features has shown good classification results both in MIR tasks [33, 35, 39, 40]) and in LID [19], we surmise that they are not as suitable for SID.

5 Summary

The analysis presented reveals tendencies concerning the application of multiple novelty beat histogram-based rhythm descriptors for SID. It has been shown that at least on one dataset of swiss german, the rhythmic features are not very helpful to achieve high accuracy in SID, although it has been shown that other rhythm metrics can capture the idiosyncrasy present in the corpus [20, 17, 41]. The reasons for that are not clear yet, but possible candidates are the specificity of the corpus language, the size of the dataset or that the features do not capture speech rhythm characteristics in a way that is speaker-specific. The latter might well be the case, as we were able to show in a previous study [19] that the same features are indeed descriptive of speech rhythm when it comes to the task of LID.

Another clue pointing to this direction is that the features achieved good accuracy for a few speakers, showing that they could partly capture characteristics of specific speakers, but not in every case. However, further tests with other datasets are necessary to confirm this tendency. From a theoretical perspective they are nevertheless very useful, as they give clues to the importance of speech rhythm for the corresponding task. The simple spectral features have shown very high performance with a low computational cost and should therefore be further applied.

Future work will include the following tasks: The use of larger datasets as the GLOBALPHONE [42] in order to be able to draw conclusions across languages and to test for rhythmic variability both between speakers and between languages at the same time. Further feature analysis is also scheduled, so as to investigate if the tendencies observed in the present study are robust across datasets and other settings (speaker, elicitation methods), as well as further investigating which aspect of the speech data (language, dataset size, feature parameterization etc) is the most important in generating better results. Specifically with respect to the speech tempo, an automatic tempo extraction scheme similar to the one used here, such as the tempogram [43], will be used in combination with manually obtained ground truth data in order to investigate the validity of the automatic tempo extraction procedure. Finally, further rhythm feature extraction algorithms, e.g. the modulation scale spectrum [44] or similarity detection schemes [45] will be adapted so as to be used for speech rhythm description.

References

[1] Ramus, F., Nespor, M., and Mehler, J., "Correlates of linguistic rhythm in the speech signal," *Cognition*, 73(3), pp. 265–292, 1999.

[2] Grabe, E. and Low, E. L., "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, 7(515-546), 2002.

[3] Dellwo, V., Fourcin, A., and Abberton, E., "Rhythmical classification of languages based on voice parameters," in *ICPhS '07*, pp. 1129–1132, 2007.

[4] Dellwo, V. and Koreman, J., "How speaker idiosyncratic is measurable speech rhythm," in *Abstract presented at the annual IAFPA meeting*, 2008.

[5] Arvaniti, A. and Ross, T., "Rhythm classes and speech perception," *Understanding Prosody: The Role of Context, Function and Communication*, 13, p. 75, 2012.

[6] Tilsen, S. and Arvaniti, A., "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages," *The Journal of the Acoustical Society of America*, 134(1), pp. 628–639, 2013.

[7] Dauer, R. M., "Stress-timing and syllable-timing reanalyzed," *Journal of phonetics*, 1983.

[8] Dellwo, V., "Rhythm and speech rate: A variation coefficient for DeltaC," *Language and language-processing*, pp. 231–241, 2006.

[9] Arvaniti, A., "The usefulness of metrics in the quantification of speech rhythm," *Journal of Phonetics*, 40(3), pp. 351–373, 2012.

[10] Arvaniti, A. and Rodriquez, T., "The role of rhythm class, speaking rate, and F0 in language discrimination," *Laboratory Phonology*, 4(1), pp. 7–38, 2013.

[11] Turk, A. and Shattuck-Hufnagel, S., "What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong," *Laboratory Phonology*, 4(1), pp. 93–118, 2013.

[12] Farinas, J., Pellegrino, F., Rouas, J.-L., and André-Obrecht, R., "Merging segmental and rhythmic features for automatic language identification," in *Audio, Speech and Signal Processing, 2002. ICASSP 2002. International Conference on*, volume 1, pp. I–753, 2002.

[13] Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R., "Modeling prosody for language identification on read and spontaneous speech," in *Acoustics, Speech and Signal Processing, 2003. ICASSP 2003. IEEE International Conference on*, volume 6, pp. I–40, 2003.

[14] Rouas, J.-L., Farinas, J., and Pellegrino, F., "Automatic modelling of rhythm and intonation for language identification," in *International Conference on Phonetic Sciences*, pp. 567–570, 2003.

[15] Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R., "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, 47(4), pp. 436–456, 2005.

[16] Rouas, J.-L., "Automatic prosodic variations modeling for language and dialect discrimination," *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(6), pp. 1904–1911, 2007.

[17] Dellwo, V., Leemann, A., and Kolly, M.-J., "Speaker idiosyncratic rhythmic features in the speech signal." in *INTERSPEECH*, 2012.

[18] Tilsen, S. and Johnson, K., "Low-frequency Fourier analysis of speech rhythm," *The Journal of the Acoustical Society of America*, 124(2), pp. EL34–EL39, 2008.

[19] Lykartsis, A. and Weinzierl, S., "Using the beat histogram for speech rhythm description and language identification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] Dellwo, V., Kolly, M.-J., and Leemann, A., "Speaker identification based on temporal information: a forensic phonetic study of speech rhythm and timing in the Zurich variety of Swiss German, International Association for Forensic Phonetics and Acoustics Conference," *Santander, Spain*, 2012.

[21] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., and Torres-Carrasquillo, P. A., "Support vector machines for speaker and language recognition," *Computer Speech & Language*, 20(2), pp. 210–229, 2006.

[22] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A., and Leek, T. R., "Phonetic speaker recognition with support vector machines," in *Advances in neural information processing systems*, p. None, 2003.

[23] Campbell, W. M., Sturim, D. E., and Reynolds, D. A., "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, 13(5), pp. 308–311, 2006.

[24] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., and Torres-Carrasquillo, P. A., "Support vector machines for speaker and language recognition," *Computer Speech & Language*, 20(2), pp. 210–229, 2006.

[25] Mandasari, M. I., McLaren, M., and van Leeuwen, D. A., "Evaluation of i-vector speaker recognition systems for forensic application." in *INTERSPEECH*, pp. 21–24, Citeseer, 2011.

[26] Matějka, P., Glembek, O., Castaldo, F., Alam, M. J., Plchot, O., Kenny, P., Burget, L., and Černocky, J., "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Acoustics, Speech and Signal Processing, 2011. ICASSP 2011. IEEE International Conference on*, pp. 4828–4831, IEEE, 2011.

[27] Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., and Vaquero, C., "Unsupervised domain adaptation for i-vector speaker recognition," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[28] Garcia-Romero, D. and McCree, A., "Supervised domain adaptation for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4047–4051, IEEE, 2014.

[29] Greenberg, C. S., Bansé, D., Doddington, G. R., Garcia-Romero, D., Godfrey, J. J., Kinnunen, T., Martin, A. F., McCree, A., Przybocki, M., and Reynolds, D. A., "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, pp. 224–230, 2014.

[30] Senior, A. and Lopez-Moreno, I., "Improving DNN speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 225–229, IEEE, 2014.

[31] Lerch, A., *An introduction to audio content analysis: Applications in signal processing and music informatics*, Wiley & Sons, 2012.

[32] Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. B., "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, 13(5), pp. 1035–1047, 2005.

[33] Tzanetakis, G. and Cook, P., "Musical genre classification of audio signals," *IEEE transactions on Speech and Audio Processing*, 10(5), pp. 293–302, 2002.

[34] Burred, J. J. and Lerch, A., "A hierarchical approach to automatic musical genre classification," in *DAFx*, 2003.

[35] Gouyon, F., Dixon, S., Pampalk, E., and Widmer, G., "Evaluating rhythmic descriptors for musical genre classification," in *AES '04*, pp. 196–204, 2004.

[36] Campbell, W. M., Singer, E., Torres-Carrasquillo, P. A., and Reynolds, D. A., "Language recognition with support vector machines," in *ODYSSEY04 – The Speaker and Language Recognition Workshop*, 2004.

[37] Vapnik, V., *The nature of statistical learning theory*, Springer, 2000.

[38] Dellwo, V., Leemann, A., and Kolly, M.-J., "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," *The Journal of the Acoustical Society of America*, 137(3), pp. 1513–1528, 2015.

[39] Lykartsis, A., Wu, C.-W., and Lerch, A., "Beat histogram features from NMF-based novelty functions for music classification," in *ISMIR*, 2015.

[40] Lykartsis, A. and Lerch, A., "Beat histogram features for rhythm-based musical genre classification using multiple novelty functions," in *DAFx*, 2015.

[41] Dellwo, V., Leemann, A., and Kolly, M.-J., "The recognition of read and spontaneous speech in local vernacular: The case of Zurich German," *Journal of Phonetics*, 48, pp. 13–28, 2015.

[42] Schultz, T., "Globalphone: a multilingual speech and text database developed at karlsruhe university." in *INTERSPEECH*, 2002.

[43] Grosche, P. and Müller, M., "Extracting predominant local pulse information from music recordings," *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(6), pp. 1688–1701, 2011.

[44] Marchand, U. and Peeters, G., "The Modulation Scale Spectrum and its Application to Rhythm-Content Description." in *DAFx*, pp. 167–172, 2014.

[45] Pohle, T., Schnitzer, D., Schedl, M., Knees, P., and Widmer, G., "On Rhythm and General Music Similarity." in *ISMIR*, pp. 525–530, 2009.