Athanasios Lykartsis, Stefan Weinzierl

Using the Beat Histogram for Speech Rhythm Description and Language Identification

Lykartsis, Athanasios; Weinzierl, Stefan (2015): Using the Beat Histogram for Speech Rhythm Description and Language Identification. In: INTERSPEECH 2015 - 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015. pp. 1007–1011. https://www.isca-speech.org/archive/interspeech_2015/i15_1007.html

Using the Beat Histogram for Speech Rhythm Description and Language Identification

Athanasios Lykartsis[1], *Stefan Weinzierl*[1]

[1]Audio Communication Group, Technische Universität Berlin, Germany

athanasios.lykartsis@tu-berlin.de, stefan.weinzierl@tu-berlin.de

Abstract

In this paper we present a novel approach for the description of speech rhythm and the extraction of rhythm-related features for automatic language identification (LID). Previous methods have extracted speech rhythm through the calculation of features based on salient elements of speech such as consonants, vowels and syllables. We present how an automatic rhythm extraction method borrowed from music information retrieval, the beat histogram, can be adapted for the analysis of speech rhythm by defining the most relevant novelty functions in the speech signal and extracting features describing their periodicities. We have evaluated those features in a rhythm-based LID task for two multilingual speech corpora using support vector machines, including feature selection methods to identify the most informative descriptors. Results suggest that the method is successful in describing speech rhythm and provides LID classification accuracy comparable to or better than that of other approaches, without the need for a preceding segmentation or annotation of the speech signal. Concerning rhythm typology, the rhythm class hypothesis in its original form seems to be only partly confirmed by our results.

Index Terms: speech rhythm, beat histogram, language identification, novelty functions, rhythm typology

1. Introduction

Speech rhythm and its quantification has been an interesting and controversial matter of research, with implications for language rhythm typology and the possible existence of two or more rhythmic classes of languages (stress- and syllable-timed) [1, 2], dubbed the *Rhythm Class Hypothesis* [3]. In the last years, analysis of speech rhythm has focused on the attempt to obtain metrics of acoustic correlates of speech rhythm which could provide information about the rhythmic patterns of speech, generally by manually annotating vowels, consonants and stresses in the speech signal and consequently calculating statistics of the durations between intervals of those language prominence units, resulting in measures such as the ΔC, $\%V$, *nPVI* and *VarcoC* [3, 4, 5, 6]. Those metrics have been proven useful as first attempts to design descriptors of speech rhythm and were very often used to investigate language rhythm typology, by testing for significant differences between languages and attempting to position the languages in a rhythm continuum between stress- and syllable-timed [4, 5]. For various small speech corpora, they have provided evidence that supports the rhythm class hypothesis and have therefore been seen as adequate measures of speech rhythm [4, 5, 6]. However, the scientific discussion about speech rhythm and its measurement continues up to the present day [7, 8, 9, 10]. In this context, the aforementioned metrics have also been criticized [9, 11, 12, 13] as not being robust with respect to the information they hold about speech rhythm, since differences between languages have not been consistent or significant across all studies, presumably due to the existence of many non-language specific factors affecting speech rhythm [9]. Other shortcomings are the manual annotation necessary for the procedure (which is tedious and can be subjective or erroneous), their derivation on basis of abstract language elements (i.e., syllables) as opposed to quantities physically manifest in the speech signal (e.g., its amplitude envelope or other measures) and, finally, their variability with respect to other, non-rhythm-related speech parameters such as speaker or elicitation method [6, 9]. However, several promising studies on the description of speech rhythm have taken a different direction, attempting to extract rhythmic quantities directly from the acoustic signal, specifically by extracting salient periodicities and their characteristics from its amplitude envelope [14, 15]. Furthermore, studies from the field of language discrimination [16, 17] have used measures derived from fundamental frequency and amplitude to discriminate between pairs of languages with relative success. Other studies from the field of rhythm- or prosody-based automatic language identification (LID) [18, 19, 20, 21, 22] have conducted rhythm modeling by using schemes such as automatic segmentation of the speech signal in pseudosyllables and extracted statistical features describing energy and fundamental frequency which produced good results (in the area of $60 - 80\%$) in LID tasks, showing that those features can indeed be useful for describing speech rhythm. The crux of those approaches is that the focus is shifted on quantities in the speech signal rather than on the regularities of more linguistically defined speech elements. This paper follows in that rationale, introducing a novel method for speech rhythm description, inspired from similar rhythm analysis methods from the field of Audio Content Analysis [23], which have been used for tasks such as musical genre classification [24] with success. We assume that the rhythmic content of a sound can be captured through the signal-inherent periodicities and their properties. This definition does not differentiate between musical and speech signals, providing a unified concept for rhythm which has been called for [25, 26]. In the following chapters, the rhythm features are described, after determining the signal properties whose periodicities are relevant for rhythm. The features are evaluated in an automatic LID task for two established multilingual speech corpora in order to draw conclusions about their suitability for rhythm-based LID and on rhythm language typology. Results are encouraging regarding the feature capacity, but with certain caveats which are discussed. Moreover, findings concerning language rhythm classes are ambivalent. Finally, advantages and disadvantages of the proposed method and the most informative features are discussed and perspectives for further research are given.

2. Method

Various approaches for rhythm description and quantification have been developed in the field of Music Information Retrieval (MIR) [27]. In the context of musical genre classification, the focus lay on the extraction of signal periodicities from a musical excerpt. Beginning with the work of Scheirer [28], a representation for periodicities of the signal amplitude envelope in the lower frequency area was introduced for beat tracking. Tzanetakis and Cook [29] modified and used this representation, called the *beat histogram*, for extracting rhythmic content features. Similar approaches followed also by Burred and Lerch [30] and Gouyon et al. [31]. The fundamental assumption is that those features are representative of the regularities in the temporal structure of an acoustic signal, describing multiple aspects of the signal's inherent periodicities. For both music and speech, the beat histogram captures periodicities related to strong, recurring 'beats', in effect salient onsets of the signal's constituent elements.

The beat histogram calculation can take place on basis of the trajectory of various relevant signal quantities over time [32]. As such, the representation will then express periodicities related to this quantity, which might have different statistical and other properties than those which are amplitude or energy related. Those temporal trajectories are called *novelty functions* [33]. A careful consultation of the most important works in phonetics, MIR and rhythm-based LID (mentioned in Section 1), as well as a study of the important rhythm definition approaches in music theory and cognition [34, 25, 35] reveals that there are three essential quantities whose temporal evolution must be taken into account for the the extraction of speech rhythm: The **amplitude** of the signal envelope is an acoustic correlate of perceived loudness. This makes it the basis for the detection of rhythm which results from the changing energy of the signal due to the application of *stresses* on specific parts of speech in comparison to others. As such, it denotes *intonation*. The **pitch** or value of a salient (for speech, the fundamental *F0*) frequency in the signal and the temporal trajectory thereof is the most important tonal rhythm carrier in the signal and expresses *speech prosody*. Features derived from its beat histogram can describe changes in voice melody trajectories, regularities in rising or falling voice pitch or related changes. **Spectral** changes are an acoustic correlate of change in sound texture and timbre, which essentially characterize different categories of sounds (such as tonal or noisy) or changes in spectral content (e.g. high or low-frequency content). Features from a beat histogram based on spectrum novelty can serve as descriptors for change of speech elements, such as consonants and vowels, or even different formants. In our study, amplitude novelty is extracted through the calculation of the **RMS amplitude** of the signal. The **fundamental frequency (F0)** is extracted through the use of a spectral harmonic product algorithm on a filtered version of the speech signal (using a 4th-order Butterworth lowpass with a 800 Hz cutoff-frequency), so as to ensure tracking of the fundamental frequency alone. Three standard features are extracted to track spectral changes [23]: the **spectral flux (SF)** (indicating general spectral change), the **spectral flatness (SFL)** (as a measure of signal tonalness or noisiness) and the **spectral centroid (SCD)** (a measure of the spectral centre-of-weight), the latter also on a filtered version of the signal (using a 4th-order Butterworth bandpass filter between 300 Hz and 3200 Hz) to ensure that only formant area frequencies are considered. More information on those features can be found in [23]. Experiments in musical genre classification using features based on similar amplitude, tonal and spectral shape novelty functions have shown promising results for a wide range of datasets [32], suggesting their suitability for LID, a task analogue to genre classification [26].

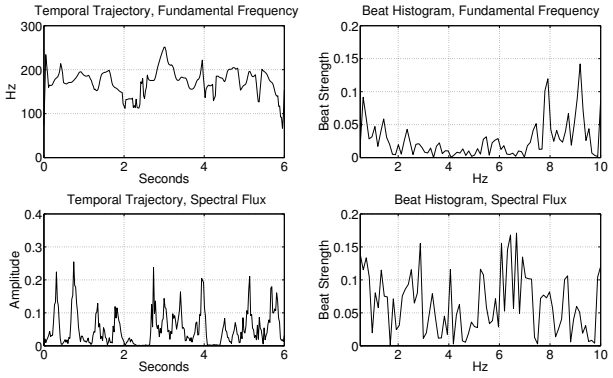

Figure 1: *Novelty functions and corresponding beat histograms.*

Distribution	Peak
Mean (ME)	Salience of Strongest Peak (A1)
Standard Deviation (SD)	Salience of 2nd Stronger Peak (A2)
Mean of Derivative (MD)	Period of Strongest Peak (P1)
SD of Derivative (SDD)	Period of 2nd Stronger Peak (P2)
Skewness (SK)	Period of Peak Centroid (P3)
Kurtosis (KU)	Ratio of A0 to A1 (RA)
Entropy (EN)	Sum (SU)
Geometrical Mean (GM)	Sum of Power (SP)
Centroid (CD)	
Flatness (FL)	
High Frequency Content (HFC)	

Table 1: *Subfeatures extracted from beat histograms.*

Fig. 1 gives an overview of two novelty functions (F0 and Spectral Flux) and their corresponding beat histograms for the utterance *Gestern war ich in einem Selbsterfahrungskurs. Ich bin mir nicht wirklich sicher, ob es mir gefallen hat* (from the german subset of the MULTEXT corpus, signal 11). It is clear that the two novelty functions show different periodicities and therefore carry valuable information on multiple speech rhythm levels. More specifically, the fundamental frequency follows the prosody of the given utterance, whereas the spectral flux measure is expected to track general changes of spectrum in the signal, i.e. phoneme or stress changes.

The beat histogram computation follows the computation in [29, 32]. All spectral-based features are calculated through a Short-Time-Fourier-Transform (STFT), whereas the fundamental frequency and the RMS measure on basis of the time-domain signal, both of which with the same temporal resolution parameters from the time domain signal. The complete procedure for the generation of a feature vector representing each utterance includes the following steps: the audio signal is downmixed to mono, resampled to 22.5 kHz, DC-freed and normalized. Afterwards, the signal is separated in texture windows with a length of 3 s and 50% overlap, on which the beat histograms are extracted. The STFT is performed with a frame-length of 46.4 ms, a Hann window and an overlap of 75%, whereas for the time-domain features the same parameters are used. The novelty function is computed through the calculation of the temporal trajectory of the features and half-wave rectification. The beat histogram is extracted through an Autocorrelation Function (ACF) for each texture window, retaining the

area between 0.5 Hz and 10 Hz, as representative for the relevant periodicities in speech [25]. Finally, the beat histograms extracted from all 3 s frames for an utterance are averaged. For each beat histogram, two categories of features can be extracted (Table 2). Similar features on beat histograms have been used in [29, 30, 31], providing valuable statistical information on the temporal features of each language, similar to the work on LID in [15]. In total, 5 novelty functions are used for the production of as many beat histograms, from each of which 19 subfeatures are extracted, producing in total 95 features.

3. Experimental Setup and Evaluation

For evaluation, extraction of a series of non-rhythmic features was undertaken, by calculating their values over all texture windows (keeping the average value inside an analysis window) on a speech file. Those non-rhythmic features serve as a baseline for the comparison, since acoustic feature-based LID-approaches are among those providing very high performance [36, 37]. Acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs) and Shifted Delta Cepstral (SDCs) features have been used widely for non-rhythmic LID with good results [38, 39]. However, for the sake of comparability with the novel rhythm features, we used a baseline set which comprises all five novelty functions which were also utilized for the calculation of the beat histograms. From their temporal trajectories, the distribution features listed in Table 2 were extracted. In total, the baseline feature set comprises 5 novelties times 11 subfeatures $=$ 55 features. For supervised classification, the Support Vector Machines (SVM) [40] algorithm under MATLAB with a Radial Basis Function (RBF) kernel in a multiclass setting was used. A grid search procedure (i.e. a search for the optimal parameter values) was applied to determine the hyperparameters for this kernel (C, γ). All experiments took place with a 10-fold cross-validation, with results averaged over the folds. Z-score standardization was conducted prior to classification, separately for the train and test set. Classification performance was evaluated through *accuracy*, defined as the proportion of correctly classified samples to all samples.

As speech material, two established multilingual speech corpora for automatic LID were used: the MULTEXT PD [41] and the OGI-MLTS corpus [42]. The first is a corpus of read, high quality speech which contains five indoeuropean languages which are assumed belong to the two basic rhythm groups (english and german to the stress-timed, french, italian and spanish to the syllable-timed), making it useful to test the rhythm class hypothesis when using the proposed novel features. The corpus 10 speakers per language (5 male and 5 female with an average of 15 passages per speaker) and an average length of 20 s for each utterance. The OGI-MLTS corpus contains spontaneous, telephone quality speech from eleven languages (featuring apart from indoeuropean also tonal languages such as mandarin chinese, or even others, such as hindi or vietnamese), multiple speakers per language (male and female) and an average length of 45 s for each utterance. For the experiments in this paper, we retained only the four languages which are common with the MULTEXT PD corpus and which can be used for rhythm typology research. The two selected datasets represent two cases of speech material with very different properties.

In order to identify the best performing descriptors and novelty functions we conducted feature selection following two approaches: First, we apply a filter method (Mutual Information with Target Data [43], using the maximum relevance CMIM metric [44] from the MI-Toolbox [45]). From the feature ranking, we retain the N best features which gave comparable accuracy to the full rhythmic feature set. Second, we evaluate each of the five novelty functions separately, by retaining only the 19 subfeatures resulting from the corresponding beat histogram.

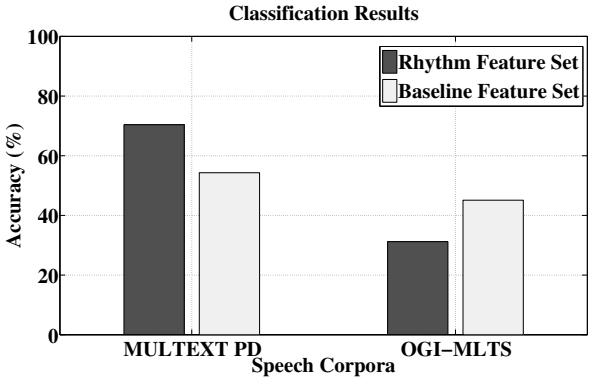

Figure 2: *Classification results, both datasets and feature sets*

True/Predicted	EN	FR	GE	IT	SP
EN	110	22	5	5	8
FR	4	76	5	9	6
GE	15	22	121	23	19
IT	3	21	5	114	7
SP	7	25	5	6	107
Acc. (%)	73.3	76	60.5	76	71.3
Prior (%)	20	13.3	26.7	20	20

Table 2: *Confusion matrix for MULTEXT PD corpus languages, rhythmic features, average accuracy:* 70.4%.

True/Predicted	EN	FR	GE	SP
EN	58	56	33	39
FR	38	76	27	45
GE	44	55	39	48
SP	44	56	27	49
Acc. (%)	31.2	40.9	21.0	26.3
Prior (%)	25	25	25	25

Table 3: *Confusion matrix for OGI-MLTS corpus languages, rhythmic features, average accuracy:* 31.2%.

Rank	MULTEXT PD	OGI-MLTS
1	FL.SFL	SP.HPS
2	GM.SF	P3.HPS
3	A2.SF	A2.HPS

Table 4: *Best features after filter feature selection. Abbreviation left of point denotes subfeature, otherwise novelty function.*

Rhythmic feature subset	MULTEXT PD	OGI-MLTS
All features	70.4	31.2
RMS Amplitude	67.5	25.4
Fundamental Pitch	70.4	27.4
Spectral Flux	67.5	25.5
Spectral Flatness	66.8	24.7
Spectral Centroid	64.9	24.9

Table 5: *Group feature selection, results given in percentages.*

4. Results

Results of the classification procedure are presented in Fig. 2. For the MULTEXT PD corpus, the rhythmic feature set per-

forms better than the baseline set. The performance of the baseline set (54.3%) lies close to that of the rhythmic feature set (70.4%). For the OGI-MLTS dataset, results show the exact opposite tendency: The baseline set with an accuracy of (37.5%) outperforms the rhythmic set (31.5%). With regards to the performance of the corpora, a great difference in accuracies can be observed: Whereas the MULTEXT PD corpus shows a satisfactory performance which lies well above the average prior (20%), the OGI-MLTS corpus accuracy stays at relatively low levels (which are, however, comparable to those in other rhythm modeling LID studies [22]). In the case of the MULTEXT PD corpus, high accuracy can be achieved for all languages. In the case of the OGI-MLTS corpus, only french shows better performance, whereas the accuracy for other languages is only moderately above the prior. The confusion matrices for both cases are given in Tables 2 and 3. It can be seen that in the case of the MULTEXT PD corpus, the rhythm class hypothesis is confirmed only partly: The hypothesized stress-timed languages english and german are not confused with each other more than with others outside this group. In the syllable-timed group, italian and spanish are confused with french, but not with each other. However, the tendency towards misclassifications toward french can be observed for all languages. For the OGI-MLTS corpus, specific misclassifications between languages in the hypothesized same rhythm class, such as english-german or french-spanish, cannot be observed in this case as well. Finally, concerning feature selection for the rhythmic feature set, results show that the same accuracy can achieved with the first 19 (MULTEXT PD) or 21 (OGI-MLTS) features of the CMIM ranking. In Table 4, a list of the best features for both datasets is given. It is noted that between the novelty functions, such based on spectral flux, spectral flatness and pitch are most commonly among the best ones. Finally, selection based on novelty feature groups (Table 5) shows that all novelty functions are almost equally important for accuracy, a result which is true for both corpora. In both cases, the F0 feature subgroup seems to perform marginally better than the others.

5. Discussion

The results presented in Section 4 suggest that the application of the beat histogram features for automatic LID is indeed valuable, since it provides comparable performance to that of other rhythm-based LID approaches [18, 19, 21, 22], although latest i-vector-based methods provide even higher results [46, 47]. The differences observed between the rhythmic and baseline feature sets are telling with respect to the robustness and quality of the proposed features. In the case of the MULTEXT PD corpus, which is a prosodic database, rhythmic features seem better suited to capture differences between languages than more general acoustic features. On the other hand, for the more generic OGI-MLTS corpus, non-rhythmic features perform better, showing that in that case rhythm features are informative enough. Other reasons which could explain the difference in performance between the two datasets are the signal quality, which in case of the OGI-MLTS corpus might impair the extraction of rhythm features or features in general significantly; and the difference in speech elicitation method, showing that spontaneous speech not only makes the extraction of robust features much more difficult, but also does not allow rhythmic features to achieve acceptable performance. Those observations are useful in determining the scope of use of the suggested rhythmic features, suggesting that they could be more suitable for read speech with good signal quality, but their robustness could be further improved. With regards to the best features, the fact that novelty functions of pitch and spectral change features produce the most salient beat histograms is a hint for their eligibility for speech rhythm analysis. It is interesting that features such as P1 and P2 (showing periodicities of prominent beats in speech) are not among the best ones. This hints towards the fact that either speech periodicities cannot help differentiate between languages (as they could be noisy because of variability due to other factors) or that they cannot be reliably extracted from the beat histograms through the subfeatures presented here. Concerning language typology on basis of the beat histogram features, the rhythm class hypothesis does not seem to be corroborated in its pure form from our results on the MULTEXT PD corpus: on the one hand it is clear that languages supposed to be rhythmically close to each other, such as english and german are not confused with each other more than with languages from different supposed rhythm classes. On the other hand, spanish and italian are more confused with french than with english or german (which would hint towards a rhythmic similarity in this group), however this can be an artifact of the specific dataset, since french seems to act as an attractor for all other languages, hinting that its rhythmic features are somehow representative of other languages as well. In the case of the OGI-MLTS corpus, results also do not confirm the rhythm class hypothesis directly. Those results can indicate that the novel features in their present form are better suited for specific languages. However, they might also be the consequence of our features not capturing speech rhythm in the same form as the rhythm class hypothesis first posited. More experiments are needed in order to determine of those results are dependent on dataset or the feature extraction and classification methods.

6. Conclusions

The presented beat histogram features are shown to be good descriptors of speech rhythm since they have been shown to provide good accuracy in an automatic LID task. Furthermore, the features achieved accuracies comparable to those of other speech rhythm feature approaches [21, 22] for the same datasets, further attesting to the merit of the method. Amongst the advantages of the presented rhythm description scheme is that it does not require any preprocessing such as syllable annotation or even automatic segmentation which is time-consuming or could potentially insert erroneous assumptions. Furthermore, the method allows the automatic processing of greater datasets and provides a novel perspective on the description of speech rhythm through solely signal-based measures. However, more experiments with greater corpora (such as GLOBALPHONE [48]), extraction parameters (to test, e.g., for effects concerning the texture window size) and other classification methods (such as artificial neural nets, as well as unsupervised methods) will be conducted, so as to be able to check for result consistency and improve robustness. Furthermore, the relationship between the features and more abstract speech elements is not entirely clear, prompting future research to establish concrete connections. Further future work on feature selection will attempt to find out which novelty functions and features are the most informative across many datasets and experimental setups, in order to compare the results with those from phonetics or human speech rhythm perception research. Concerning language typology, the presented rhythm-based LID does not seem to corroborate the rhythm class hypothesis in its pure form, but gives incentives to attempt and reformulate the hypothesis in a new version so as to account for the empirical evidence.

7. References

[1] K. L. Pike, *The Intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

[2] D. Abercrombie, *Elements of general phonetics*. Edinburgh University Press Edinburgh, 1967, vol. 203.

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

[4] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[5] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

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

[7] P. Wagner, "The rhythm of language and speech: Constraining factors, models, metrics and applications," *Germany: Habilitationsschrift, University of Bonn*, 2008.

[8] L. Wiget, L. White, B. Schuppler, I. Grenon, O. Rauch, and S. L. Mattys, "How stable are acoustic metrics of contrastive speech rhythm?" *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1559–1569, 2010.

[9] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *Journal of Phonetics*, vol. 40, no. 3, pp. 351–373, 2012.

[10] A. Turk and S. Shattuck-Hufnagel, "What is speech rhythm? a commentary on arvaniti and rodriquez, krivokapić, and goswami and leong," *Laboratory Phonology*, vol. 4, no. 1, pp. 93–118, 2013.

[11] P. Roach, "On the distinction between stress-timed and syllable-timed languages," *Linguistic controversies*, pp. 73–79, 1982.

[12] W. J. Barry, B. Andreeva, M. Russo, S. Dimitrova, T. Kostadinova *et al.*, "Do rhythm measures tell us anything about language type," in *ICPhS '03*, 2003, pp. 2693–2696.

[13] A. Arvaniti, "Rhythm, timing and the timing of rhythm," *Phonetica*, vol. 66, no. 1-2, pp. 46–63, 2009.

[14] S. Tilsen and K. Johnson, "Low-frequency fourier analysis of speech rhythm," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. EL34–EL39, 2008.

[15] S. Tilsen and A. Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 628–639, 2013.

[16] F. Cummins, F. Gers, J. Schmidhuber, and C. Elvezia, "Automatic discrimination among languages based on prosody alone," *Speech Communication*, 1999.

[17] A. Thymé-Gobbel and S. E. Hutchins, "Prosodic features in automatic language identification reflect language typology," in *ICPhS '99*, 1999.

[18] J. Farinas, F. Pellegrino, J.-L. Rouas, and R. André-Obrecht, "Merging segmental and rhythmic features for automatic language identification," in *ICASSP '02*, vol. 1, 2002, pp. I–753.

[19] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Modeling prosody for language identification on read and spontaneous speech," in *ICASSP '03*, vol. 6, 2003, pp. I–40.

[20] J.-L. Rouas, J. Farinas, and F. Pellegrino, "Automatic modelling of rhythm and intonation for language identification," in *ICPhS '03*, 2003, pp. 567–570.

[21] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, no. 4, pp. 436–456, 2005.

[22] J.-L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1904–1911, 2007.

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

[24] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, March 2006.

[25] A. D. Patel, *Music, language, and the brain*. Oxford university press, 2008.

[26] S. Hübler and R. Hoffmann, "Comparing the rhythmical characteristics of speech and music–theoretical and practical issues," in *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*. Springer, 2011, pp. 376–386.

[27] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Computer music journal*, vol. 29, no. 1, pp. 34–35, 2005.

[28] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[29] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[30] J. J. Burred and A. Lerch, "A hierarchical approach to automatic musical genre classification," in *DAFX '03*, 2003.

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

[32] A. Lykartsis, "Evaluation of accent-based rhythmic descriptors for genre classification of musical signals," Master's thesis, Audio Communication Group, Technische Universität Berlin, 2014.

[33] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[34] F. Lerdahl and R. S. Jackendoff, *A generative theory of tonal music*. MIT press, 1983.

[35] J. London, *Hearing in time*. Oxford University Press, 2012.

[36] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *Signal Processing Magazine, IEEE*, vol. 11, no. 4, pp. 33–41, 1994.

[37] M. A. Zissman and K. M. Berkling, "Automatic language identification," *Speech Communication*, vol. 35, no. 1, pp. 115–124, 2001.

[38] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification." in *INTERSPEECH*, 2003.

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

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

[41] E. Campione and J. Véronis, "A multilingual prosodic database." in *ICSLP*, vol. 98, 1998, pp. 3163–3166.

[42] Y. K. Muthusamy, R. A. Cole, B. T. Oshika, L. D. Consortium *et al.*, "The ogi multi-language telephone speech corpus," in *ICSLP*, vol. 92, 1992, pp. 895–898.

[43] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[44] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[45] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[46] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 271–284, 2007.

[47] M. Li and W. Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," *submitted to INTERSPEECH*, 2014.

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