Analysis of Speech Rhythm for Language Identification Based on Beat Histograms

Athanasios Lykartsis[1], Alexander Lerch[2], Stefan Weinzierl[3]

[1] *Audio Communication Group, TU Berlin, 10587 Berlin, Germany, Email: athanasios.lykartsis@tu-berlin.de*

[2] *Georgia Institute of Technology, Center for Music Technology, GA 30332 Atlanta, USA, Email: alexander.lerch@gatech.edu*

[3] *Audio Communication Group, TU Berlin, 10587 Berlin, Germany, Email: stefan.weinzierl@tu-berlin.de*

Introduction

Rhythm is a basic property of acoustic signals [1][2], with a presumed common basis for its perception grounded both in speech and music [3], hinting towards a similarity which can be tracked in the acoustic signals as well. For speech signals, rhythm analysis can provide relevant conclusions both with respect to linguistic questions (e.g. language rhythm typology) and for applications in speech technology (e.g. in multilingual dialogue systems). However, speech rhythm is difficult to analyze, since its modeling or measurement are not straightforward.

In phonetics, the measurement of speech rhythm has mainly been performed by the development of statistical measures (known as *rhythm metrics*) that capture the patterns of intervals of and between salient speech elements such as vowels, consonants and syllables. Such metrics include the standard deviation of consonant intervals ΔC, the percentage of vocalic intervals %V and the Pairwise Variability Index (PVI) [4][5][6]. Although they have been used extensively for speech rhythm description and the investigation of rhythmical differences between languages, those measures have also been criticized [7] for lack of robustness and for producing inconsistent results with respect to the rhythm class hypothesis, which states that languages belong either to a stress-timed or to a syllable-timed group [8]. Further problems include the manual or automatic annotation of speech elements which is required in order to perform the analysis, as well as that the focus lies only on high-level language elements (such as syllables or consonants-vowels) and their duration patterns for rhythm description instead of examining directly measurable signal properties. Various technical attempts to model rhythm were also undertaken in the field of rhythm-based language identification (LID). A number of studies ([10][11][12][13]) have extracted rhythmic units by using the concept of automatic segmentation in pseudosyllables (structures of the form C^nV, where C is a consonant and V a vowel) and calculating parameters concerning duration and properties of speech elements such as fundamental frequency or energy. Such studies have achieved satisfactory results ($60 - 80\%$) in rhythm-based LID for a number of speech corpora, which shows the importance of rhythm and prosody based features for the LID task. They still, however, bear the disadvantage of taking into account higher-level language units such as syllables to extract speech rhythm.

In order to overcome these problems, we propose an alternative approach for rhythm extraction and modeling for LID. We draw inspiration from the field of Music Information Retrieval (MIR), where there have been numerous approaches for rhythm extraction, for instance for the problem of automatic musical genre classification. One of the widely used representations is the *Beat Histogram*, which has emerged as a method for rhythmic content description for audio classification and has been described in [15][16][17]. Its basic premise is that the rhythm of an audio excerpt can be described through creating a representation of the distribution of its periodicities in a very low frequency area and extracting relevant statistical and other properties from it. A similar approach has been recently presented by Tilsen & Arvaniti [9], who modelled speech rhythm by extracting periodicities and from the signal envelope and analyzing their relationships.

This paper describes the use of the beat histogram for the creation of speech rhythm features for LID by using several relevant signal properties as the basis for its creation. The goals are the evaluation of those novel features for rhythm-based LID and the analysis of speech rhythm through investigation of the rhythm class hypothesis. In the following, the methods for speech rhythm feature extraction are described. The classification setup with two supervised learning algorithms as well as the experimental results for one multilingual speech corpus are presented and discussed.

Method

The beat histogram is created through the extraction of the temporal trajectory of a given signal quantity or its difference (also known as a Novelty Function [18]). After the signal is preprocessed (mean removal, filtering etc.), the novelty function of the signal amplitude or its envelope is calculated, half-wave rectified and periodicities are represented for an area typically between 30 and 300 BPM, by using a method such as the Autocorrelation Function (ACF) [15], the Discrete Fourier Transform (DFT) or the comparison with a filter bank of tuned bandpass filters [14][16]. The end result is a compact representation of the magnitude and value of all important signal periodicities, where for example the tempo (main periodicity) of the analyzed track can be observed. The properties of the rhythmic content of the excerpt can be then extracted with the use of descriptors such as the mean, standard deviation and other distribution statistics, as well as more specific descriptors such as the amplitude and frequency of the most salient peak.

In the context of rhythm description and musical genre classification, most of the studies have used the beat histogram with the signal amplitude envelope as a novelty function. This approach, however, does not take into ac-

count changes in other signal properties such as tonal or general spectral changes which might have other periodicities. Therefore, it is sensible to expand the beat histogram by taking into account novelty functions of other signal properties whose change over time is relevant. Experiments in musical genre classification using amplitude, tonal and spectral shape novelty functions have shown promising results for a wide range of datasets [19]. This approach is therefore adapted here for speech: We use three categories of novelty functions so as to capture the characteristics of the most important temporal trajectories in the signal:

- **amplitude-based**, accounting for changes in signal energy or loudness reflecting changing intonation,
- **fundamental frequency-based (F0)**, tracking changes in speech prosody and
- **spectral shape-based**, accounting for changes in spectral content which reflect changes of speech elements (consonants/vowels) or phoneme position.

The F0 is extracted through the use of a harmonic product spectrum algorithm [20] on a filtered version of the speech signal (4th-order Butterworth lowpass with a cutoff frequency at 800 Hz) so as to ensure tracking of the fundamental frequency alone. Three established features are extracted to track spectral changes: the **spectral flux** (indicating general spectral change), the **spectral flatness** (indicating tonalness/noisiness) and the **spectral centroid** (a measure of the spectral centre-of-weight), the latter also on a filtered version of the signal (4th-order Butterworth bandpass between 300 Hz and 3300 Hz) to insure that only formant area frequencies are considered. From the corresponding beat histograms, we then extract a list of standard features, relating to periodicity distribution statistics and to the position and salience of the beat histogram peaks, which can be seen in table 1. More information on the extracted novelty functions as well as on the subfeatures listed in Table 1 can be found in [21]. For the beat histograms here, a periodicity range from 0.5 Hz to 10 Hz was selected as characteristic of the most important speech rhythms. All features are extracted over a frame-by-frame basis: The speech files are separated in short frames of 3 s with an overlap of 50%. In the case of the rhythmic features, the beat histograms are averaged over all frames and the subfeatures extracted from them. In total, the rhythm feature set comprises 5 features times 19 subfeatures = 95 features.

Table 1: Subfeatures extracted from Beat Histograms.

Distribution	Peak
Mean (ME)	Salience of Strongest Peak (A1)
Standard Deviation (SD)	Salience of 2nd Stronger Peak (A0)
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)	

Experiments and Results

Experiments

In order to be able to conduct an evaluation of the proposed beat histogram features, a baseline feature set needed to be established. To that purpose, extraction of a series of non-rhythmic features was undertaken, by calculating the feature values over all texture frames (by keeping the average value inside an analysis window) of a speech file. A number of acoustic features such as MFCCs, LPCs and SDCs have been used widely for non-rhythmic language identification [22][23]. In order to maximize comparability with our rhythmic features and to be able to estimate the merit introduced by the use of the beat histograms, we used as a baseline feature set all five novelty functions listed in the previous section. The features on each novelty function can be seen in the *Distribution* column of Table 1. In total, the baseline feature set comprises 5 features times 11 subfeatures = 55 features. For supervised classification, we use the established Support Vector Machines (SVM) [24], which have been used extensively and has shown good results in many classification problems up to date. For comparison, we also use the basic and simple k-Nearest-Neighbors algorithm. For the SVM algorithm the Radial Basis Function (RBF) Kernel is used with the parameters C and γ determined through grid search, while for the kNN algorithm the euclidean distance was used with $k = 1, 3, 5$. All experiments take place as multiclass one-vs-one classification problems with 10-fold cross validation and standardization of the features (z-score, separately for train and test set). In order to evaluate the classification we use the average *accuracy* (Acc.) as a performance measure, defined as the proportion of correctly classified samples to all samples classified, which can be easily derived from the confusion matrices as the sum of the diagonal to the total samples count.
We tested our features on one established multilingual speech corpus, MULTEXT [25]. This is a read speech dataset, comprising five indoeuropean languages (English, French, German, Italian and Spanish) with high signal quality (20 kHz sample rate, 16 bit quantization depth). The dataset contains between 10 and 20 passages with an average length of 20 s from 10 speakers per language (5 male and 5 female). The choice of this dataset is of importance, since it has been used extensively for rhythm-based LID and can allow conclusions both to rhythm-based automatic LID performance and language typology, since the languages contained are those prototypically belonging to the two basic groups after rhythm class hypothesis [8]: English and German to the stress-timed, French, Spanish and Italian to the syllable-timed.

Results

Results of classification can be seen in Fig. 1 and Tables 2 and 3. Concerning classification accuracy, two tendencies can be observed: First, concerning overall accuracy, the SVM algorithm outperforms the kNN in all cases, with the kNN showing very low scores (even for $k = 5$ which was the best case). Second, for the SVM, the rhythmic

feature set has slightly better accuracy than the baseline, whereas for the kNN, the baseline set shows moderately better performance than the rhythmic set. Furthermore, for the SVM results are clearly above the average prior (Pr.) of 20%(the percentage of the samples of each class in the dataset) and satisfactory, whereas for the kNN, accuracy is low and below the prior for both feature sets. With regards to language rhythm typology, the pure form of the rhythm class hypothesis does not seem to be confirmed in either case: for the kNN, all languages are confused with English, except for English itself which is classified as German, hinting towards a rhythmic similarity only between stress-timed languages. For the SVM, all languages are confused with French, but neither Italian and Spanish nor German and English are confused with each other more than with other languages.

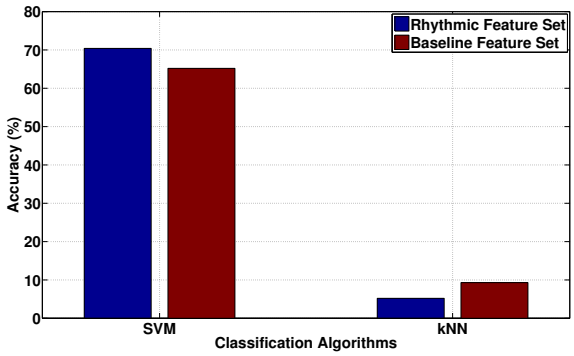

Figure 1: Classification results

Table 2: Confusion Matrix for the kNN algorithm, rhythmic feature set. All numbers indicate sample count, rows Acc. and Pr. are given as percentages, average accuracy 7.2%

True	Predicted				
	Eng.	Fre.	Ger.	Ita.	Spa.
Eng.	**12**	27	66	24	21
Fre.	62	**2**	33	2	1
Ger.	121	23	**40**	12	4
Ita.	97	14	38	**0**	1
Spa.	90	15	42	0	**1**
Acc.	**8**	**2**	**20**	**0**	**1**
Pr.	**20**	**13.7**	**27.3**	**20**	**20**

Table 3: Confusion Matrix for the SVM algorithm, rhythmic feature set. All numbers indicate sample count, rows Acc. and Pr. are given as percentages, average accuracy 70.4%

True	Predicted				
	Eng.	Fre.	Ger.	Ita.	Spa.
Eng.	**110**	22	5	5	8
Fre.	4	**76**	5	9	6
Ger.	15	22	**121**	23	19
Ita.	3	21	5	**114**	7
Spa.	7	25	5	6	**107**
Acc.	**73.3**	**76**	**60.5**	**76**	**71.3**
Pr.	**20**	**13.7**	**27.3**	**20**	**20**

Discussion

The results presented in the previous Section are promising for further research: it is clear that the use of beat histogram features can be useful for rhythm-based LID: The identification accuracy using the SVM algorithm (70.4%) lies in the same range or is better than the ones achieved in other studies [13] ($67 \pm 8\%$). An interesting point is the performance of the baseline set, which is comparable to that of the rhythmic feature set, showing that the rhythmic features can explain as much language-specific variability in the speech signal as simple, more general features. However, it should be noted that the use of other general features can achieve even higher performance scores [22][23], indicating that there is room for improvement, e.g. through the use of other novelty functions or features on the beat histogram.
With respect to classification algorithms, the SVMs are definitely advantageous in performance for the rhythmic feature set. Indeed, the very low performance of the kNN (in contrast to the comparable performance for the baseline set) is an indication that rhythmic features require more robust machine learning algorithms for identification, a result which has been confirmed through other studies [11]. Another possible reason for the lower performance of the kNN is the lack of a sufficient number of training samples or the relatively high number of features, resulting in the curse of dimensionality [21]. A possible amendment would be to perform a principal component analysis (PCA) and use the most important components as features.
Concerning the speech corpus itself, the attained performance shows that for read speech of good quality, features can be extracted which are informative of the rhythmic content and can be used to identify languages on that basis. With respect to the different languages of this dataset, it is interesting to observe from the confusion matrix (Table 2) how French seems to act as a "universal attractor" for all other classes. This effect could be due to actual difference of the french language rhythm in comparison to other languages, or to particular characteristics of this specific speech corpus (such as it containing spontaneous speech). In general, the rhythm class hypothesis, which would classify English and German together as stress-timed and the other as syllable-timed languages, does not seem to be corroborated on basis of those data.

Conclusions

In this paper we presented first results on the use of novel features for rhythm analysis and rhythm-based LID. The use of the beat histogram for speech rhythm analysis is innovative and results are promising, harboring their further use. For the rhythm descriptors, not only the signal amplitude but also other rhythm-relevant signal quantities were used as basis for the creation of the beat histogram. Furthermore, a comprehensive array of subfeatures was extracted from the beat histograms, which provides ample information about the periodicities in the signal and their patterns. We could show that classifica-

tion performance for one multilingual speech corpus using the SVM algorithm is comparable to that of similar studies and close to that when using other basic, non-rhythmic features. The proposed method has the advantage that it takes into account the rhythmic on the signal and not on the speech element level, which throws a new light on speech rhythm and allows its analysis from different aspects. Another important advantage of the proposed method for speech rhythm analysis is that it is fully automatic and can be extended for larger datasets, while providing significant information on speech periodicities. This provides another aim for further research: The application of the method to speech corpora with different content (such as the OGI-MLTS [26], which contains more languages and spontaneous speech) and which are much more comprehensive (such as the GLOBALPHONE [27]) is scheduled. At this point, the relation of the rhythm features to other speech rhythm metrics and language elements such as syllables and consonant-vowel clusters is unclear, suggesting another direction for future work. Further future goals include the investigation of optimal parameter settings for feature extraction and the conduct of feature selection to identify the most informative features, as well as the utilization of unsupervised classification methods with the focus on evaluating the method and clarifying its merits for rhythm-based LID.

References

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

[2] London, J.: Hearing in time. Oxford University Press, New York, 2012.

[3] Hübler, S., Hoffmann, R.:. Comparing the rhythmical characteristics of speech and music – Theoretical and practical issues. Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues (2011), 376-386.

[4] Ramus, F., Nespor, M., Mehler, J.: Correlates of linguistic rhythm in speech signal. Cognition 73.3 (1999), 265-292.

[5] Grabe, E., Low, E. L.: Durational variability in speech and the rhythm class hypothesis. Papers in Laboratory Phonology 7. Cambridge University Press, Cambridge, 2002.

[6] Dellwo, V.: Rhythm and speech rate: A variation coefficient for ΔC. Language and language-processing (2006), 231-241.

[7] Arvaniti, A.: The usefulness of metrics in the quantification of speech rhythm. Journal of Phonetics 40.3 (2012), 351-373.

[8] Abercrombie, D. (ed.): Elements of General Phonetics. Edinburgh University Press, Edinburgh, 1967.

[9] Tilsen, S., Amalia 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 (2013), 628-639.

[10] Farinas, J., Pellegrino, F.: Automatic rhythm modeling for language identification. INTERSPEECH (2001), 2539-2542.

[11] Pellegrino, F. Chauchat, J.-H., Rakotomalala, R., Farinas, J.: Can automatically extracted rhythmic units discriminate among languages?, Speech Prosody (2002).

[12] Rouas, J.-L., Farinas, J., Pellegrino, F.: Automatic modelling of rhythm and intonation for language identification. 15th ICPHS (2003), 567-570.

[13] Rouas, Jean-Luc, Farinas, J., Pellegrino, F., André-Obrecht, R.: Rhythmic unit extraction and modelling for automatic language identification. Speech Communication 47.4 (2005), 436-456.

[14] Scheirer, E. D.: Tempo and beat analysis of acoustic musical signals. The Journal of the Acoustical Society of America 103.1 (1998), 588-601.

[15] Tzanetakis, G., Cook, P.: Music genre classification of audio signals. IEEE Transactions on Speech and Audio Processing 10.5 (2002), 293-302.

[16] Burred, J. J., Lerch, A.: A hierarchical approach to automatic musical genre classification. Proceedings of the 6th Int. Conference on Digital Audio Effects (2003), 8-11.

[17] Gouyon, F., Dixon, S., Pampalk, E., Widmer, G.: Evaluating rhythmic descriptors for musical genre classification. Proceedings of the 25th International AES Conference (2004), 196-204.

[18] Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M. B.: A tutorial on onset detection in music signals. IEEE Trans. on Speech and Audio Processing 13.5 (2005), 1035-1047.

[19] Lykartsis, A.: Evaluation of Accent-Based Rhythmic Descriptors for Genre Classification of Musical Signals. Master's Thesis, Technische Universität Berlin, Berlin, 2014.

[20] Noll, M.: Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate. Proceedings of the Symposium on Computer Processing in Communications 19 (1970), 779-797.

[21] Lerch, A.: An introduction to Audio Content Analysis, Wiley, New York, 2012.

[22] Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D. A.: Acoustic, phonetic, and discriminative approaches to automatic language identification, INTERSPEECH (2003).

[23] Campbell, W. M., Singer, E., Torres-Carrasquillo, P.A., and Reynolds, D. A.: Language recognition with support vector machines, ODYSSEY04 (2004).

[24] Vapnik, V. N.: Statistical Learning Theory. Wiley, New York, 1998.

[25] Campione, E., Véronis, J.: A multilingual prosodic database. ICSLP (1998), 3163-3166.

[26] Muthusamy, Y. K., Cole, R. A., Oshika, B. T.: The OGI multi-language telephone speech corpus. ICSLP (1992), 895-898.

[27] Schultz, T.: Globalphone: a multilingual speech and text database developed at karlsruhe university. INTERSPEECH (2002).