Modified Group Delay Feature for Musical Instrument Recognition

Aleksandr Diment[1], Padmanabhan Rajan[2], Toni Heittola[1], and Tuomas Virtanen[1⋆]

[1] Tampere University of Technology
aleksandr.diment@tut.fi
[2] University of Eastern Finland

Abstract. In this work, the modified group delay feature (MODGDF) is proposed for pitched musical instrument recognition. Conventionally, the spectrum-related features used in instrument recognition take into account merely the magnitude information, whereas the phase is often overlooked due to the complications related to its interpretation. However, there is often additional information concealed in the phase, which could be beneficial for recognition. The MODGDF is a method of incorporating phase information, which lacks of the issues related to phase unwrapping. Having shown its applicability for speech-related problems, it is now explored in terms of musical instrument recognition. The evaluation is performed on separate note recordings in various instrument sets, and combined with the conventional mel-frequency cepstral coefficients (MFCCs), MODGDF shows the noteworthy absolute accuracy gains of up to 5.1% compared to the baseline MFCCs case.

Keywords: Musical instrument recognition, music information retrieval, modified group delay feature, phase spectrum

1 Introduction

Situationally tailored playlisting, personalised radio and social music applications are just several examples of application of music information retrieval. This research area, broadly speaking, studies the methods of obtaining information of various kinds from music.

Musical instrument recognition is one example of its subtopics, and it has been most actively explored since the 1990's. An extended overview of the early systems is given in [7]. The recent works on the subject propose novel classification approaches in terms of the given problem (e.g., genetic algorithms [15] and semi-supervised learning [4]), as well as introduce new features (e.g., multiscale MFCCs [18]). There exists an established set of features commonly applied for instrument recognition. Depending on whether they treat audio from the temporal or spectral point of view, these are subcategorised accordingly.

⋆ This research has been funded by the Academy of Finland, project numbers 258708, 253120 and 265024.

The temporal features (e.g., the amplitude envelope) address instrument recognition under the assumption that the relevant information is within the transient properties of a signal. Such assumption is perceptually motivated: the attack characteristics are believed to play crucial role in human recognition of musical instruments [5].

The spectral features employ a different approach. Particularly, those that are related to the harmonic properties of a sound (e.g., inharmonicity and harmonic energy skewness) do preserve the important properties of the musical instrument timbre [1]. The same applies to other spectrum-related features as well, such as mel-frequency cepstral coefficients (MFCCs). It is worth mentioning that, being spectrum-based features, they in fact concentrate only on its magnitude part.

In general, spectral information is complete only if both magnitude and phase spectra are specified. Signal processing difficulties, such as wrapping of the phase, make direct processing of the phase spectra challenging. A popular solution is to use the *modified group delay function* [17], which can be applied to process information from the phase spectrum. Previously, it has been utilised for several applications, including speech recognition [17, 11] and spectrum estimation [19].

The *modified group delay feature (MODGDF)* has not yet been applied for instrument recognition, although phase information has been recently incorporated in the neighbouring areas (e.g., instrument onset detection by means of the phase slope function [12]). This work proposes calculating MODGDF for pitched instrument recognition, either primarily or as a complement to the established MFCCs, under the assumption that phase may contain additional information relevant in terms of instrument classification. The primary objective is to demonstrate whether MODGDF is at all capable of introducing improvement in the performance of an instrument recogniser.

This paper is organised as follows. Section 2 presents the motivation and properties of MODGDF. Subsequently, Section 3 introduces the particular instrument recognition system, which incorporates MODGDF, as well as MFCCs. Its performance is consecutively evaluated with various instrument sets in Section 4. Finally, the conclusions about the applicability of the feature are drawn along with the future research suggestions in Section 5.

2 Modified Group Delay Feature

This section commences with the motivation behind utilising phase information in general and MODGDF in particular for musical instrument recognition. This is followed by the details of computation of the group delay function, as well as its modified version along with the reasons for that modification.

2.1 Motivation

Phase is often overlooked in many audio processing solutions due to the complications related to the unwrapping of the phase spectrum. Despite of that, phase could be highly informative due to its ability of indicating peaks in the spectral

envelope. In terms of speech recognition and related problems, these correspond to formants, which are useful for extracting speech content.

In the musical instrument signals, however, the presence of formants in the spectrum is not as strong [13], or they are not a factor independent from fundamental frequency, in contrast to speech signals. For example, in the spectra of trombone or clarinet, due to the acoustical change of active volume of their body during the sound production, the resonances depend on pitch [8, 14].

Nevertheless, a phase-based feature appears to be applicable for instrument recognition as well. Broadly speaking, while the commonly applied MFCCs feature is capable of modelling the resonances introduced by the filter of the instrument body, it neglects the spectral characteristics of the vibrating source, which also play their role in human perception of musical sounds [9]. Incorporating phase information attempts to preserve this neglected component.

Additionally, a phase-related feature could indicate a so-called *mode locking* phenomenon, characteristic to some instruments. The frequencies of each mode are never in precise integer ratios, which would yield a nonrepeating waveform. However, despite the inharmonicieties of the natural resonances, the individual modes of such instruments are locked into the precise frequency and phase relationships, provided certain conditions are met [8].

2.2 Properties

The Fourier transform $X(\omega)$ of a signal $x[n]$ in the polar form is expressed as

$$X(\omega) = |X(\omega)| \exp^{j\theta(\omega)} . \tag{1}$$

The *group delay function* is obtained as [2]

$$\tau_g(\omega) = -\mathrm{Im}\left(\frac{d}{d\omega}\log(X(\omega))\right) \tag{2}$$

$$= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{3}$$

where $Y(\omega)$ is the Fourier transform of $y[n]$, and $y[n] = nx[n]$. The advantage of Equation 3 is that no explicit phase unwrapping is needed.

The group delay function is well-behaved only if the zeros of the system transfer function are not close to the unit circle. The zeros may be introduced by the excitation source or as a result of short time processing [3, 11]. When zeros of the transfer function are close to the unit circle, the magnitude spectrum exhibits dips at the corresponding frequency bins. Due to this, the denominator term in Equation 3 tends to a small value, resulting in a large value of the group delay function $\tau_g(\omega)$. This manifests itself in spurious high amplitude spikes at these frequencies, masking out the resonance structure in the group delay function.

The modification [17] of the group delay function is performed by suppressing the zeros of the transfer function. This is done by replacing the magnitude

spectrum $X(\omega)$ by its cepstrally smoothed version $S(\omega)$. Additionally, the parameters α and γ are introduced to control the dynamic range. The modified function is defined as

$$\tau_m(\omega) = \text{sign}\big(\tau(\omega)\big)\ \big(|\tau(\omega)|\big)^{\alpha}, \tag{4}$$

where

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}}. \tag{5}$$

In the latter equation, $S(\omega)$ is the cepstrally smoothed version of $X(\omega)$, and the function "sign" returns the sign.

To convert the function into features, the DCT is applied on Equation 4. This performs decorrelation, and the the first N_C coefficients are retained (excluding the zeroth coefficient). The parameters α, γ and the length of the cepstral lifter window lifter_w are tuned to a specific environment. In practice, the feature is calculated in short frames under the assumption of spectral stationariness within their length, and the Fourier analysis is performed with the aid of DFT.

3 System Description

The details of the developed musical instrument recognition system that incorporates MODGDF as one of its features are addressed in this section. The upcoming paragraphs are following the implementation of its building blocks.

3.1 Feature Extraction

As primarily explored features, MODGDF, as well as its first and second derivatives, are incorporated in the calculation. Additionally, a baseline scenario is included, i.e., the calculation of the static and delta MFCCs. Those are currently quite commonly applied for musical instrument recognition, proven to be amongst the most effective features [6] due to their ability to parametrise the rough shape of the spectrum, which is different for each instrument. The mel transformation, which is included in the calculation of MFCCs, is based on human perception experiments and has been demonstrated to effectively represent perceptually important information in terms of instrument recognition [16]. The classification results produced by recognisers based on MFCCs have been shown to resemble human classifications in terms of the similar confusions [6]. This baseline MFCCs scenario is intended to indicate the expected performance of the system with the given data when utilising such established feature.

Frame-blocking is performed with 50%-overlapping windows of length 20 ms. The number of mel filters in MFCCs calculation is 40. For both features, the number of extracted coefficients (referred to as N_C in the case of MODGDF) is set to 16. The search for the optimal value of the additional parameters of MODGDF (α, γ and lifter_w) is omitted due to the high computational requirements of such operation, as well as motivated by the fact that the ultimate objective is to

demonstrate whether MODGDF is at all capable of introducing improvement. The values of these parameters are therefore set to $\gamma = 0.9$, $\alpha = 0.4$ and $\text{lifter}_w = 8$, shown to be optimal for speech recognition [11]. To search for the refined values appears to be an appealing problem left for the future exploration.

The calculation of the combination of these features is foreseen in order to investigate whether the MODGDF, if not as effective as MFCCs *per se*, is capable of enhancing the performance of the system when used as a complement to the baseline feature. This way, a feature set that incorporates both amplitude and phase information is acquired. The combined features of dimension 64 are obtained by concatenating the values of MFCCs and MODGDF.

3.2 Training and Recognition

The training and recognition phases are performed by employing Gaussian mixture models (GMM). For each class, the feature vectors from the training data are used to train the GMM, i.e., to estimate the parameters of such model that best explains these features. The expectation-maximisation (EM) algorithm is used for this purpose, and each class is represented by a GMM of 16 components.

In recognition, the trained models of each class are fit into each frame of the test instances, producing log-likelihoods. The latter are summed over the frames of the test instance. Thereupon, the label of the class whose model has produced the highest log-likelihood is assigned to that instance.

4 Evaluation

The performance of the proposed approach is evaluated in a separate note-wise instrument classification scenario. Several instrument sets grouped by the level of complexity of the resulting problem are considered. The instrument content of these sets is presented below, followed by the obtained evaluation results.

4.1 Acoustic Material

The recordings (sampling frequency 44.1 kHz) used in evaluation originate from the RWC Music Database [10]. Each of the instruments is represented in most cases by three instances, which stand for different instrument manufacturers and musicians. These are subdivided into subsets according to the playing styles (e.g., bowed vs plucked strings), and only one playing style per instrument is taken into account. In total, five instruments sets (Table 1) are considered: three generic (consisting of 4, 9 and 22 various instruments) and two specific ("woodwinds" and "strings"). The choice of instruments in the sets "4 various" and "9 various" is influenced by the requirement of a sufficiently high number of notes per instrument for their consistent representation in the database. The "22 various" set, consisting of diverse instruments and even vocals, not necessary sufficiently represented in the database, is intended to demonstrate a highly complex classification scenario. The possible performance improvements with this set would ultimately indicate the real-life applicability of the proposed method.

Table 1. Instrument sets used in evaluation.

Instrument set	List of instruments
4 various	Acoustic Guitar, Electric Guitar, Tuba, Bassoon
9 various	Piano, Acoustic Guitar, Electric Guitar, Electric Bass, Trombone, Tuba, Bassoon, Clarinet, Banjo.
22 various	"4 various" + "9 various" + "woodwinds" + "strings" + vocals: Soprano, Alto, Tenor, Baritone, Bass
woodwinds	Oboe, Clarinet, Piccolo, Flute, Recorder
strings	Violin, Viola, Cello, Contrabass

Additionally, the sets "woodwinds" and "strings" are used in evaluation in order to observe a possible dependency between the physics of the instruments and the particularities of the phase content of their sound (Section 2.1), which could manifest themselves in changes in recognition accuracy.

The dataset, where each instrument is represented by several hundred recordings, is divided into the training and testing subsets. The subsets are acquired from different instrument instances in order to resemble a real-life application scenario. The ratio between the sizes of the training and test subsets is roughly 70%/30%. Due to a limited representation of some instruments, somewhat unstable results could be expected. Hence, the evaluation is performed three times with randomisation over the dataset contents, and the results are averaged.

4.2 Results

The evaluation results obtained with each of the instrument sets are summarised in Table 2. One may observe that MODGDF, used as such, is capable of serving as a reliable feature in several cases (namely, the "4 various", "woodwinds" and "strings" sets). In the case of the "strings" set, it even manages to noticeably outperform the MFCCs by 4.1%.

Considering the combined case, including MODGDF as an additional feature clearly improves the recognition accuracy with each of the instrument sets. The most significant improvement is observed with the "4 various" set, which is most

Table 2. Evaluation results.

Instrument set	Recognition accuracy, %		
	MFCCs	MODGDF	combined
4 various	90.9	84.4	96.0
9 various	82.6	59.9	84.9
22 various	68.8	41.7	70.7
woodwinds	74.5	66.7	77.2
strings	69.7	73.8	73.6

extensively represented (about 750 notes per instrument). With the more complicated sets, e.g., "woodwinds", whose representation has been not as high (about 250 notes per instrument), the baseline performance and the improvement introduced by incorporating MODGDF are rather low. In the case of "strings", the combined method appears to yield slightly lower recognition accuracy, compared to the purely MODGDF, however, such small difference is rather to be caused by the randomisation effects within the EM algorithm. Most importantly, the combined method does outperform the baseline in this case as well.

A somewhat more specific comparison of the features can be performed by observing the instrument-wise accuracies. As seen in Figure 1, obtained with the "9 various" set, most of the improvement in the recognition accuracy introduced by incorporating MODGDF is present in the cases of some of the woodwinds (Bassoon) and brass instruments (Tuba). This shows the potential of the applicability of the feature to these instrument groups, which is to be discovered in future experiments and analysis with relation to the physics of the instruments.

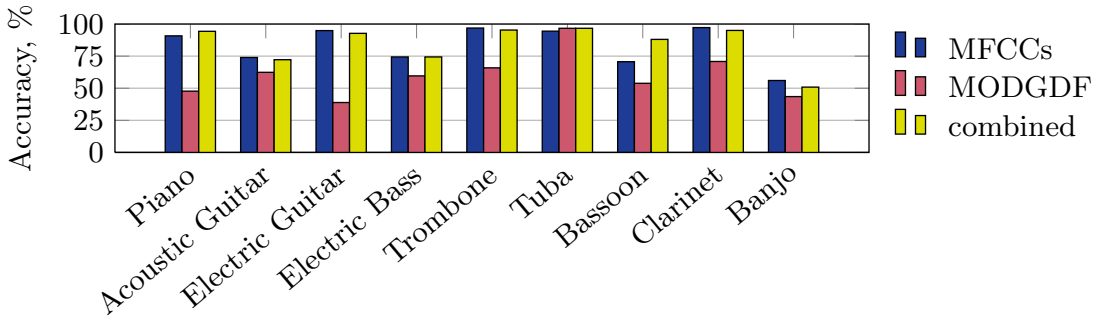

Fig. 1. Instrument-wise accuracies in each of the evaluation scenario with the "9 various" instrument set.

5 Conclusions

The proposed method of utilising MODGDF as a complementary feature to MFCCs for musical instrument recognition has yielded increase in the recognition accuracy with each of the instrument sets compared to the purely MFCCs case. The value of the accuracy increase has been shown to be up to a rather noteworthy 5.1%.

As the future suggestions, it is worthwhile to study the dependency between the physics of particular instruments and the performance of MODGDF. Additionally, a search for the optimal parameters of MODGF in terms of the given problem would be rather beneficial. Finally, evaluating the performance of the combined features after applying dimensionality reduction and decorrelation appears reasonable.

References

1. G. Agostini, M. Longari, and E. Pollastri. Musical instrument timbres classification with spectral features. In *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, pages 97–102, 2001.
2. H. Banno, J. Lu, S. Nakamura, K. Shikano, and H. Kawahara. Efficient representation of short-time phase based on group delay. In *Proc. Int. Conf. Acoust. Speech Signal Process.*, volume 2, pages 861–864, 1998.
3. B. Bozkurt, L. Couvreur, and T. Dutoit. Chirp group delay analysis of speech signals. *Speech Commun.*, 49:159–176, 2007.
4. A. Diment, T. Heittola, and T. Virtanen. Semi-supervised learning for musical instrument recognition. In *21st European Signal Processing Conference 2013 (EUSIPCO 2013)*, Marrakech, Morocco, Sept. 2013.
5. C. Duxbury, M. Davies, and M. Sandler. Separation of transient information in musical audio using multiresolution analysis techniques. In *Proc. of the DAFx Conf.*, 2001.
6. A. Eronen. Comparison of features for musical instrument recognition. In *Proc. of the IEEE Workshop on Applications of Sign. Process. to Audio and Acoust.*, 2001.
7. A. Eronen. *Signal Processing Methods for Audio Classification and Music Content Analysis*. PhD thesis, Tampere University of Technology, Finland, June 2009.
8. N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer, 1998.
9. F. Fuhrmann. *Automatic musical instrument recognition from polyphonic music audio signals*. PhD thesis, Universitat Pompeu Fabra, 2012.
10. M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. In *Proc. of the 4th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 229–230, 2003.
11. R. Hegde, H. Murthy, and V. Gadde. Significance of the modified group delay feature in speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):190–202, 2007.
12. A. Holzapfel, Y. Stylianou, A. Gedik, and B. Bozkurt. Three dimensions of pitched instrument onset detection. *Audio, Speech, and Language Processing, IEEE Trans. on*, 18(6):1517–1527, 2010.
13. K. Jensen. *Timbre Models of Musical Sounds: From the Model of One Sound to the Model of One Instrument*. Report. Københavns Universitet, 1999.
14. B. Kostek and A. Czyzewski. Representing musical instrument sounds for their automatic classification. *J. Audio Eng. Soc.*, 49(9):768–785, 2001.
15. R. Loughran, J. Walker, and M. O'Neill. An exploration of genetic algorithms for efficient musical instrument identification. In *Signals and Systems Conf. (ISSC 2009), IET Irish*, pages 1–6. IET, 2009.
16. J. Marques and P. J. Moreno. A study of musical instrument classification using gaussian mixture models and support vector machines. *Cambridge Research Laboratory Technical Report Series CRL*, 4, 1999.
17. H. Murthy and V. Gadde. The modified group delay function and its application to phoneme recognition. In *Acoustics, Speech, and Signal Processing, 2003. Proc. (ICASSP '03). 2003 IEEE Int. Conf. on*, volume 1, pages I–68–71 vol.1, 2003.
18. B. Sturm, M. Morvidone, and L. Daudet. Musical instrument identification using multiscale mel-frequency cepstral coefficients. *Proc. of the European Signal Processing Conference (EUSIPCO)*, pages 477–481, 2010.
19. B. Yegnanarayana and H. Murthy. Significance of group delay functions in spectrum estimation. *Signal Processing, IEEE Transactions on*, 40(9):2281–2289, 1992.