



Audio Engineering Society Convention Paper

Presented at the 124th Convention
2008 May 17–20 Amsterdam, The Netherlands

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

A Fast Feature Extraction System On Compressed Audio Data

Tobias Friedrich¹, Matthias Grühne¹, and Gerald Schuller¹

¹*Fraunhofer IDMT, Ilmenau, Germany*

Correspondence should be addressed to Tobias Friedrich (Tobias.Friedrich@idmt.fraunhofer.de)

ABSTRACT

We describe an efficient system, which directly extracts features from compressed audio material. It consists of a time/frequency conversion method and a feature extraction algorithm. The conversion method provides the feature extraction algorithm with a suitable complex spectral representation directly from the compressed domain. It further allows to trade-off between computational complexity and conversion accuracy. Several operating points using different conversion accuracies were tested with an MPEG audio identification system in order to evaluate the identification confidence. Based on these results it is possible to reduce the computational complexity from $O(N \log N)$ to $O(N)$ compared to the conventional approach (complete decoding followed by a frequency analysis).

1. INTRODUCTION

Since the last two decades, digital content has experienced a significant proliferation. To keep track of this growth of media data, efficient non-text based search methods become increasingly important. Depending on the type of user, different scenarios demand for different solutions. For instance, the automatic identification of audio titles is gaining relevance, since the content owners, who would like to know where and how often their music is played, are confronted with an increasing variety of distribution

channels. Some of those are radio stations, internet radio or file download portals. Moreover, private users may want to get information about a track played in the radio, or organize and clean up their own music library. These scenarios demonstrate the necessity of reducing human intervention with regard to meta data generation and making content management an intuitive task. The comprehensive MPEG-7 standard [1] contains specifications for audiovisual low-level feature extraction for multimedia search and retrieval tasks. Low-level features are

for instance, audio spectrum envelope or linear prediction coefficients. However, in this paper we concentrate only on MPEG-7 compliant audio feature extraction, which is the basis for any audio identification application. Conventionally, the features are extracted from the uncompressed audio signal and used for searching tasks in audio libraries, for automatic metadata extraction as genre recognition, tempo estimation and several more. Nevertheless, audio libraries often consist of compressed audio files like MPEG1/2 Layer 3 (MP3) and MPEG-2/4 AAC (AAC). Thus, it is obvious to take the short cut and obtain audio features directly from the compressed files in order to extract the features much faster and hence to be able to achieve a faster audio identification.

2. GOALS

Our goal is to design a feature extraction system which directly operates on compressed audio data, especially on MP3 and AAC files. The resulting system should feature a reduced complexity compared to the conventional method of fully decoding and then applying an FFT for the feature analysis. Moreover, the system is desired to be compatible to existing feature databases using MPEG-7 features.

3. PROBLEMS

The main problem of the direct feature extraction approach is that the inherent time/frequency representation in the MPEG audio coders differs from the time/frequency resolution used for MPEG-7 features. Specifically, the MPEG coders generate a time/frequency description based on MDCT filter banks (for MP3 additionally a QMF filter bank) with real valued subband signals. In contrast, MPEG-7 features are based on a Short Time Fourier Transform (STFT) having a different number of subbands and complex valued subband signals. Another problem is for instance, that a feature like Zero Crossing Rate is based on the time domain description of the audio signal and hence cannot be extracted. Fortunately, features derived from the frequency domain are sufficient to guarantee a high identification rate.

4. PREVIOUS APPROACHES

As previously mentioned, the conventional approach to obtain MPEG-7 features from compressed audio data is to decode it first and then to generate the

MPEG-7 features based on the decoded time signal. But especially when searching large libraries of compressed audio files this approach can become computationally very expensive. Several works deal with the conversion between subband domain representations, especially in the field of image and video coding. In [2, 3] the conversion between different sizes of DCT transforms is given, having the drawback that they are restricted to non-lapped transforms. The patent in [4] proposes a conversion method between the MDCT and the DFT domain. It is restricted to MDCT and DFT and therewith not suitable for our purposes, since we want to include also hybrid filter banks, an integral part of MP3. The architecture presented in [5] is not restricted to the type of filter banks used. Unfortunately, the number of subbands of the different filter banks have to be multiples of each other and this is again unsuitable for our needs. However, this paper serves as the basis for a general conversion method proposed in [6], which can be applied to any maximally-decimated filter bank without condition on their sizes. Here, a conversion matrix is generated by multiplying the analysis with a synthesis filter bank. Principally, we do the same in our new approach, though, we use a universal mathematical description, the polyphase description introduced in [7]. Additionally, we extend the method by applying it to arbitrary resolution translations between synthesis and analysis filter banks in a practical way. We further adjust it to MP3 and AAC, and exploit some special properties of the so-called conversion matrix which is explained in the next section. In [8] the problem of generating a complex from a real valued spectral representation is picked up from the reverse side. Therein it is said that a desired frequency response can be approximated by means of a linear combination with constant weighting factors. This approach only allows a coarse approximation, nonetheless, having a very small computational complexity load. This approach gave the inspiration for the issue termed as spectral approximation (see section 5.2). A completely different approach is worth mentioning here which works directly on the compressed domain. It uses the MDCT coefficients as the basis for the low-level feature extraction [9]. Since there is no conversion into the DFT domain applied, this approach is restricted to the time/frequency resolution provided by the used codec. It is hence not compatible to

existing MPEG-7 feature databases.

5. NEW APPROACH

We designed a conversion system which directly converts the given time/frequency representations of MPEG audio codecs into the time/frequency representation needed for MPEG-7 features. This is realized by using special conversion matrices, whereas the time/frequency resolution as well as the accuracy of the resulting spectrum is freely scalable. To derive these matrices, we chose a mathematical approach in the z -domain, the polyphase description, addressed in [7]. Using it, we can describe e.g. the MP3 decoder filter banks (MDCT, QMF) and the STFT (Short Time Fourier Transform) for the MPEG-7 feature extraction with polyphase matrices. The polyphase description allows us to derive conversion matrices, which directly convert the given time/frequency representation into the desired one. The resulting complex spectral coefficients are then fed to the feature extraction algorithm. Additionally, those matrices are suitable for a fast implementation, since the most significant values are evenly spread along the main diagonal, whereas they decrease quickly the further we move away from it. This property admits to approximate a desired spectral representation by only calculating the strongest diagonals while omitting the multiplications with the less important ones. The complexity of this approach is closely related to the accuracy needed for the identification of an audio file. To determine this accuracy, we show identification results of tests performed on a large audio library with different levels of conversion accuracy in section 6.2.2. These tests further show that the audio feature extraction system can deal with very coarse spectral approximations. Exploiting this property leads, in general, to a reduction of the computational complexity of the conversion from $O(N \log N)$ to $O(N)$, compared to the conventional approach. Our feature extraction system targets the MP3 and AAC formats, but is generally applicable to any subband coder. However, there is one significant difference between the mentioned two formats. MP3 uses a hybrid filter bank comprising a 32-band QMF and a switchable 6/18-band MDCT, whereas AAC only uses an MDCT which can be switched between 128 and 1024 bands.

5.1. Conversion System

An overview of the conversion system is presented

in Fig. 1, where the upper half shows the conventional transcoding scheme and the lower half our direct conversion approach. In conventional transcoding, the time signal $\hat{\mathbf{X}}(z)$ is first reconstructed and then transformed into the targeted frequency representation. This intermediate step of calculating the time signal is avoided by our direct conversion approach, where the conversion matrix $\mathbf{T}(z)$ is the matrix product of the polyphase matrices of the decoder synthesis filter bank $\mathbf{G}(z)$ and the analysis filter bank $\mathbf{H}(z)$ of the feature extraction system. A more detailed mathematical description can be found in [10].

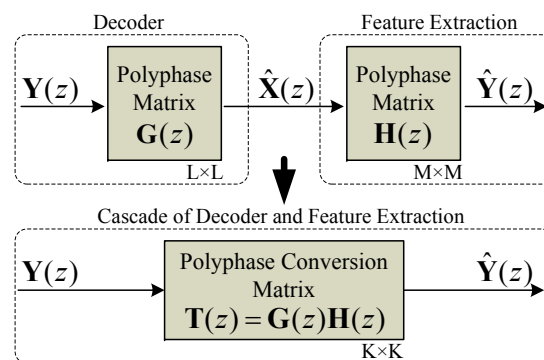


Fig. 1: Block diagram of the conventional transcoding and of our direct conversion method.

5.2. Spectral Approximation

The most important characteristic of a conversion matrix $\mathbf{T}(z)$ is that it exhibits a strong similarity to diagonal and therefore sparse matrices. For instance, Fig. 2 shows an example of such a polyphase conversion matrix, where the white areas corresponds to zeros in the matrix. Observe that we have three images of matrices, because each corresponds to the coefficients of a different power of z of the polyphase matrix. The analysis time window is set to 30 ms because it is suitable for many tasks of music information retrieval. The sampling frequency is chosen to be 44.1 kHz (generally it is arbitrary), hence the matrix generates 1024 complex Fourier coefficients as output, whereas it takes 576 (the content of one MP3 granule) real valued input samples.

It can be seen in Fig. 2 that the most significant values are evenly spread along the main diagonal. If we only keep the coefficients necessary for our de-

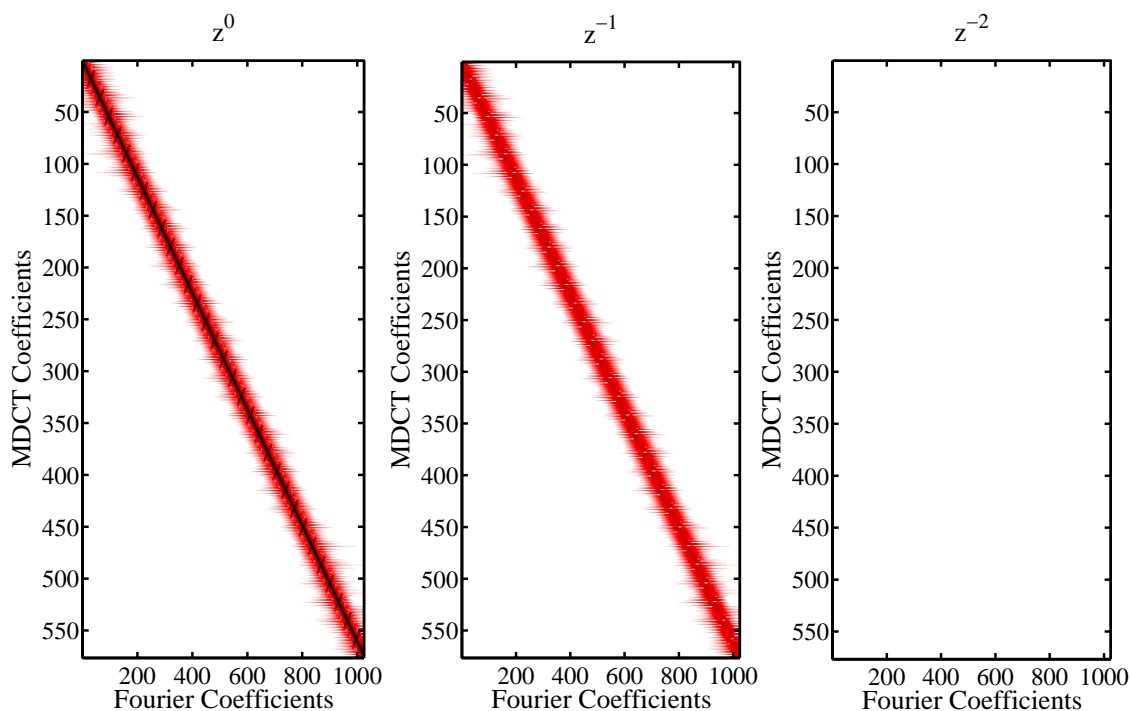


Fig. 2: Exemplary complex polyphase conversion matrix for MP3 converting one granule of 576 real valued subbands into 1024 DFT coefficients. The figure only shows absolute values.

sired accuracy, we obtain the sparse matrix shown in Fig. 3. For clarification, Fig. 4 shows an exemplaric STFT spectrum and its approximation using sparse matrices for direct conversion. For this example a conversion complexity of about 0.07% in contrast to a fully populated matrix was used.

6. PERFORMANCE EVALUATION

6.1. Complexity

In this paragraph we assess the computational complexity of the conventional MP3 decoding process that is followed by an FFT and compare it to our new approach in more detail. We include standard methods as well as highly optimized ones. Table 1 lists the number of multiplications and additions for the conventional approach.

For the standard method, the number of calculations are obtained by postulating full complexity. For instance, a standard 32-DFT needs $32^2 = 1024$ multiplications and $32^2 - 32 = 992$ additions. The data for figuring the computational amount of the optimized variants are extracted from three different papers.

	standard		optimized	
	×	+	×	+
18-IMDCT	648	612	81	149
32-QMF Synthesis	2048	1984	80	209
32-DFT/SRFFT	1024	992	68	388
Total	3720	3588	229	746

Table 1: Complexity comparison for conventionally decoding and analyzing an MP3 using standard and optimized filter bank implementations. The data of the optimized variants are extracted from [11, 12, 13].

In [12], Konstantinides found a fast implementation for QMF filter banks based on Lee’s fast DCT algorithm [14]. Marovich then reuses these results in [11] to reduce the complexity of the IMDCT. For estimating the number of calculation steps of an optimized DFT, a highly performant Split Radix Fast Fourier Transform (SRFFT) algorithm was chosen [13]. Thus, by using optimized implementations for decoding and analyzing an MP3 file the multiplica-

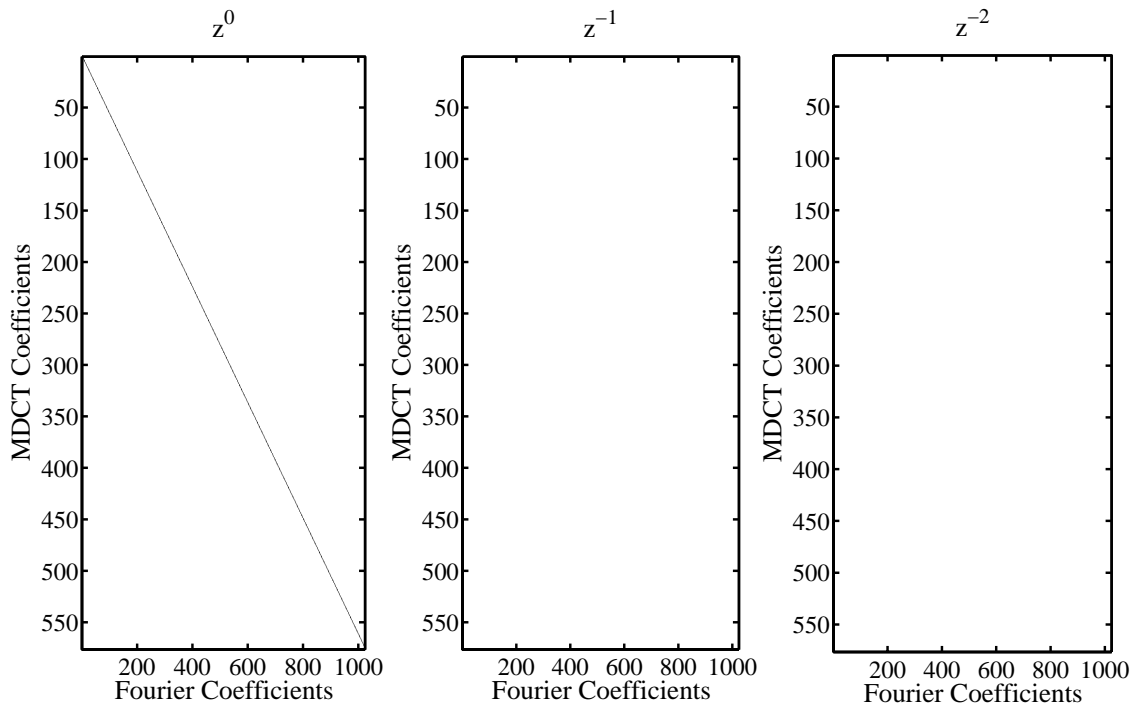


Fig. 3: Sparse polyphase matrix obtained from the conversion matrix shown in Fig. 2. Only the biggest diagonal values are maintained.

tions can be reduced by a factor of about 16 (additions by 5), compared to the standard method.

Table 2 lists the numbers of multiplications and additions for a 30 ms analysis window applied to MP3 files sampled with 44.1 kHz. The numbers of the conventional method are based on those from Table 1. The term *dense matrix* means that the matrix is fully populated so that each entry is nonzero and contributes to the computational load. The term *sparse matrix* however, stands for a conversion matrix consisting of merely the essential entries which guarantee a 100% identification confidence. Later in section 6.2.2 we investigate why it is sufficient in this case to only preserve 0.018% of the dense matrix entries. The number of additions is higher than those of multiplications, due to a fixed internal buffer updating step which consumes 2048 additions. Altogether, with our new approach we achieve a complexity reduction to a fourth of the multiplications compared to the optimized conventional one. Comparing the number of additions, we even improve by a factor of ten.

		Complexity [%]
MP3 Decoding	other MP3 blocks	19.2
	Inverse Huffman	14.4
	IQ	19.4
	Antialiasing	1.2
	IMDCT	30.8
	QMF Synthesis	15
	Total	100
FE Analysis	FFT	≈ 15

Table 3: Complexity analysis of an MP3 decoder implementation taken from [15] plus an assumed complexity of a downstream FFT. FE is an abbreviation of Feature Extraction.

A complexity analysis of a highly optimized MP3 decoder was performed by the Fraunhofer IIS in [15]. Table 3 illustrates the basic processing units of an MP3 decoder and reveals their computational amount in percent. The row indicated with *other MP3 blocks* comprises bit stream parsing as well as scalefactor, and stereo processing. *IQ* signifies In-

	Conventional				Direct Conversion			
	Standard		Optimized		Dense Matrix		Sparse Matrix	
	×	+	×	+	×	+	×	+
IMDCT	26791	25302	3349	6160				
QMF Synthesis	84672	82026	3308	8641	8128512	8126464	1463	5558
2048-DFT/FFT	4194304	4192256	16388	61444				
Total	4305767	4299584	23045	76245	8128512	8126464	1463	5558

Table 2: Complexity comparison of the conventional and our new direct conversion method for MP3 sampled at 44.1 kHz using an analysis time window of 30 ms.

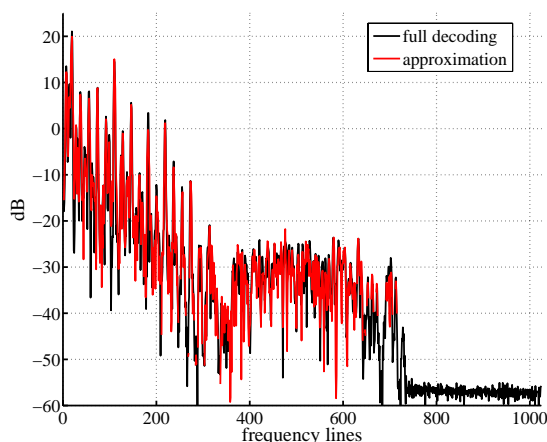


Fig. 4: Short Time Fourier Spectrum of 30 ms and its approximation using sparse matrices for direct conversion.

verse Quantization and is, together with the units mentioned before and Inverse Huffman decoding, mandatory for both, the conventional approach and our direct conversion. For the downstream FFT of the feature extraction analysis we assume an FFT to have a similar complexity as the optimized versions of an IMDCT and a QMF. The translated amount of processing steps including the FFT results to circa 44% overall. Thus, the remaining 56% comprising the filter bank operations and antialiasing, can be reduced by using our new direct conversion approach. This means, the complexity cannot be reduced more than a half. In general two different parameters describe the performance of the audio information retrieval system. One is the conversion complexity (and accuracy), the other one is the identification rate. If the conversion complexity is lowered, the approximation accuracy is reduced and accordingly

the identification rate may drop off. Depending on whether we like to have a more precise approximation of the Fourier spectrum or a computational efficient one, we can use more or fewer conversion matrix entries for the calculation.

6.2. Identification Performance

6.2.1. Test Environment

In this section we describe our experimental results of using our direct conversion method for a feature extraction for Music Information Retrieval. In order to evaluate the system, our direct conversion method was used as the input for an audio identification system. The identification system is, among others, described in [16]. It includes an extraction of a set of features, and a classification step. The classification step compares the extracted set of features against a local feature database, and retrieves the index of the song with the most similar set of features. The set of features consists of spectral flatness measures and additionally a spectrum envelope descriptor with some postprocessing methods as described in [17]. The outputs of the identification system are an index which denotes the identified song, and a value we call confidence, indicating the reliability of the result. The confidence is a heuristic of the system and is given in percent. In our experience a confidence above 50% indicates a correctly identified song. In order to assess the minimal accuracy of the direct conversion necessary to correctly identify every song, an evaluation comprising two different tests was prepared. Furthermore, two different test sets were established:

- 100 arbitrary musical test items from different genres were selected for performing a fast test, in order to find the minimum complexity thresh-

old needed which still guarantees good recognition performance.

- 775 files from 10 different genres and 63 sub-genres were chosen, in order to prove that recognition from different genres and a large database of music is possible.

In both test sets, all musical items were available in the MP3 format. The classification step performs a search on a feature database containing about 10000 items. The above mentioned test sets are a subset of this database.

The tests are:

1. The compressed files from the first test set were directly converted into the FFT domain by using a number of different conversion matrices with varying computational complexity. Altogether 20 different conversion matrices were tested.
2. From the results of the first test, a suitable conversion matrix was selected, which recognized all of the 100 files successfully and needed the least computational complexity. The needed complexity turned out to be 0.018%. For all songs of the second test set we used this conversion matrix to directly extract their feature sets and perform the classification on them.

6.2.2. Results

This paragraph describes the results of the first test shown in Fig. 5. In the figure, three different lines are depicted, indicating the minimum, maximum and average recognition confidence against the complexity. The traditional approach reaches 100% confidence on each of the items. The top line shows, that some items are already identified with confidence 100% at a complexity of 0.002%. Since we would like to reach the same classification performance as the conventional system, we have to select our direct conversion matrix according to the minimum curve. At a conversion complexity of 0.01% the minimum confidence amounts to approximately 65% and reaches 100% at a conversion complexity of 0.016%. To be on the safe side, we decided to use the direct conversion matrix with a complexity of 0.018%.

We then used this conversion matrix for the second larger test set of 775 files. The result was that all 775 files reached a confidence of 100% indicating that they were classified correctly.

7. CONCLUSIONS

A direct conversion of a time/frequency representation in compressed audio data into a time/frequency representation needed for MPEG-7 features is feasible. We found the required accuracy for this conversion is low enough for a very efficient implementation of this direct implementation, reducing the computational complexity from $O(N \log N)$ to $O(N)$.

8. ACKNOWLEDGEMENTS

This work has been partly supported by the PHAROS and the DIVAS projects, funded under the EC IST 6th Framework program. Furthermore, the work on this publication is supported by grant No. 01MQ07017 of the German THESEUS program.

9. REFERENCES

- [1] ISO/IEC, *ISO/IEC 15938-4 (MPEG-7 Audio)*, ISO, 2002.
- [2] A. N. Skodras, "Direct transform to transform computation," *IEEE Signal Processing Letters*, vol. 6, pp. 202–204, 1999.
- [3] J. B. Lee and A. Eleftheriadis, "2-D transform-domain resolution translation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 704–714, 2000.
- [4] M. M. Goodwin, "Efficient system and method for converting between different transform-domain signal representations," *US Patent 2003/0093282*, 2005.
- [5] R. K. Sande and B. Anantharaman, "An efficient VLSI/FPGA architecture for combining an analysis filterbank following a synthesis filterbank," *IEEE ISCAS*, vol. 3, pp. 517–520, 2004.
- [6] A. B. Touimi and A. Mouhssine, "Efficient conversion method between subband domain representations," *IEEE ICME*, 2005.
- [7] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Inc., 1993.

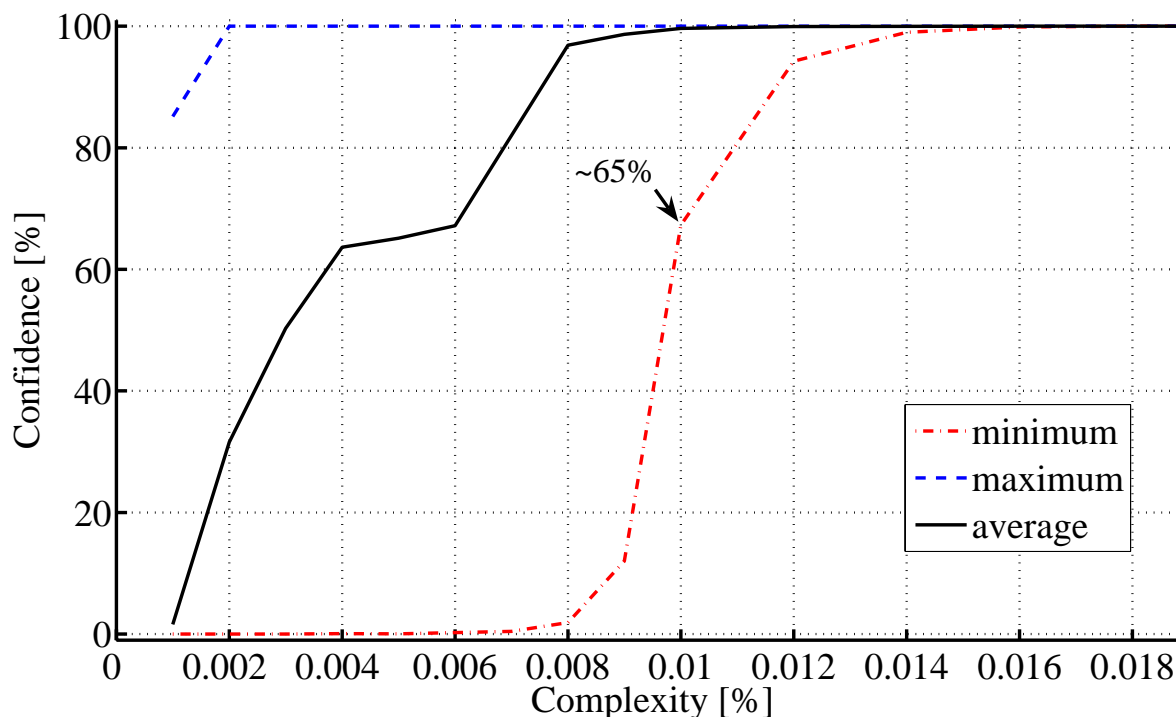


Fig. 5: Results after classification, confidence vs. conversion complexity for the test set of 100 items.

- [8] B. Edler and S. Geyersberger, "Arrangement and method for the generation of a complex spectral representation of a time-discrete signal," *EU Patent 2003/0766165*, 2004.
- [9] X. Shao, C. Xu, Y. Wang, and M. S. Kankanhalli, "Automatic music summarization in compressed domain," *IEEE ICASSP*, vol. 4, pp. 261–264, May 2004.
- [10] T. Friedrich, M. Gruhne, and G. Schuller, "Sub-band conversion for feature extraction from compressed audio," *IEEE ICASSP, Las Vegas*, April 2008.
- [11] S. B. Marovich, "Faster MPEG-1 layer III audio decoding," Tech. Rep., HP Laboratories Palo Alto, June 2000.
- [12] K. Konstantinides, "Fast subband filtering in MPEG audio coding," *IEEE Signal Processing Letters*, vol. 1, pp. 26–28, February 1994.
- [13] S. Bouguezal, M. O. Ahmad, and M. N. S. Swamy, "Arithmetic complexity of the split-radix FFT algorithms," *ICASSP*, vol. 5, pp. 137–140, March 2005.
- [14] B. G. Lee, "A new algorithm to compute the discrete cosine transform," *IEEE ICASSP*, pp. 1243 – 1245, 1984.
- [15] F. Mayer, D. Dalquen, and T. Dettbarn, "Hardware accelerating audio coding algorithms," *IEEE ICCE*, pp. 283 – 284, 2006.
- [16] E. Allamanche, J. Herre, O. Hellmuth, B. Froba, T. Kasten, and M. Cremer, "Content-based identification of audio material using MPEG-7 low level description," 2001.
- [17] E. Allamanche, J. Herre, O. Hellmuth, T. Kastner, and M. Cremer, "Apparatus and method for robust classification of audio signals," 2004, United States Patent US 10/931,635.