Bitrate and Tandem Detection for the AMR-WB Codec with Application to Network Testing

Tobias Hübschen
Digital Signal Processing and System Theory
Kiel University
Kiel, Germany
thu@tf.uni-kiel.de

Gerhard Schmidt
Digital Signal Processing and System Theory
Kiel University
Kiel, Germany
gus@tf.uni-kiel.de

***Abstract*—In network testing, identifying the cause for an observed speech quality degradation is of special interest. Common speech codec related causes to be identified are the application of a low bitrate or the occurrence of transcoding or self-tandem. This paper presents two comprehensible types of signal features which enable a speech-quality-motivated bitrate detection for the AMR-WB codec. The first type of feature is based on codec linearity, while the second type exploits the different structure of the fixed codebook at each bitrate. With these underlying features, the bitrate detection is performed with high accuracy. Since the one feature gathers information on the last applied bitrate and the other on coding effects accumulated during the entire transmission, this paper, additionally, provides a method to extract information on the occurrence of self-tandem in the network-under-test.**

***Index Terms*—network testing, AMR-WB, bitrate, self-tandem, listening quality**

I. Introduction

When conducting speech communication network tests with regard to listening quality, reference speech signals are transmitted over a network-under-test and the received signals are recorded. The signal pairs of reference and recorded signal are then fed to algorithms like [1], [2], which instrumentally determine a quality score. While it may be concluded from these scores that speech quality is impaired, it is, however, not directly possible to track the impairment back to specific elements or properties of the network. Recently, efforts have been focused on decomposing these quality scores into quality dimensions [3] or technical causes [4] to allow for a more detailed analysis of communication networks. In accordance with that, this work proposes a method which is able to track a speech quality impairment back to speech coding effects.

The majority of the recent work available in this direction focuses on the non-intrusive (without reference) codec detection for forensic purposes [5]–[7]. Although the applied speech codec does correlate with the perceived speech quality, codec type detection is not of particular interest for e.g. mobile network testing, since the speech codec is usually fixed for the network type under test. However, since the relevant speech codecs are adaptive regarding their bitrate and, therefore, speech quality, identifying the bitrate does provide useful network and quality information. Especially for the *Adaptive Multi-Rate Wideband* (*AMR-WB*) codec [8], satisfactory methods providing this information still need to be developed. Consequently, this work partly aims at developing a reliable method for detecting the applied bitrate of the *AMR-WB* codec.

Due to the context of network testing, it is possible to develop a method which relies on the availability of the reference signal. Therefore, additional coding effects may be detected which cannot be directly identified from the recorded signal. These are mainly effects caused by codec self-tandeming (multiple applications of the same codec) or transcoding within the network, which were proven to cause a significant speech quality degradation [9]. Hence, this work provides a method to both detect the bitrate and the presence of self-tandem scenarios for the *AMR-WB* codec. A high level block diagram of the system to be discussed is given in Fig. 1.

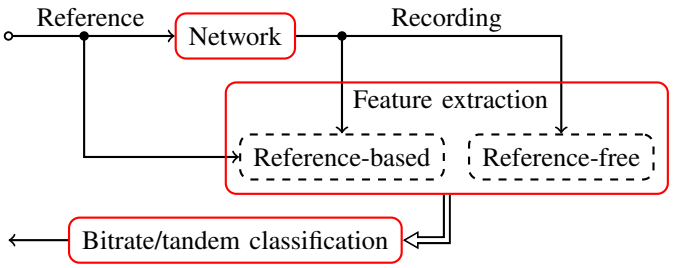

Fig. 1. High level block diagram of the proposed system.

This paper is structured as follows: Section II discusses a reference-based feature, which is targeted at capturing the characteristics of the entire transmission path. Section III then focuses on the recorded signal only to provide a computationally efficient indicator for the last used bitrate. These two features are combined in Section IV and V to build classifiers for the bitrate and self-tandem scenarios, respectively.

II. Codec Linearity

From a system-theoretical point of view, speech codecs may be described as timevariant nonlinear systems [10] in the discrete-time domain, where a linear part is described separately from a purely nonlinear part (Fig. 2). The contribution of the linear transmission to the output signal $y(k)$ is

Funded by the German Research Foundation.

ISBN 978-90-827970-1-5

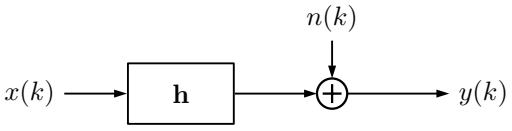

Fig. 2. System model of a speech codec.

modeled by the convolution of the input $x(k)$ with the impulse response $\mathbf{h}$ of the codec. Any other contribution to $y(k)$ which cannot be expressed through this convolution is modeled by the signal $n(k)$, which is simply added to the signal resulting from the described convolution. Consequently, $n(k)$ accounts for the purely nonlinear transmission of the codec. Since the correlation between a linearly and a nonlinearly transmitted signal of same origin is generally low, $n(k)$ may be modeled as additive noise.

Based on this model, measures for the degree of linearity/nonlinearity of a speech codec were applied for the detection of two codec classes, where the class assignment was in accordance with the codecs' bitrate [10], [11]. While these methods were used to distinguish between different narrowband codecs, it is modified in this work to distinguish between the bitrates of a multi-rate wideband codec (*AMR-WB*).

Assuming the absence of correlation between the original speech signal $x(k)$ and its nonlinearly transmitted counterpart $n(k)$, an estimate of the linear transfer function may be computed by

$$\hat{H}(\mu) = \frac{\hat{\Phi}_{xy}(\mu)}{\hat{\Phi}_{xx}(\mu)}, \tag{1}$$

where μ is the frequency index of the *Discrete-Fourier-Transform* and $\hat{\Phi}_{xx}(\mu)$ and $\hat{\Phi}_{xy}(\mu)$ are the estimates of the auto and cross power spectral densities, respectively. The two densities are separately estimated using [12]. For the computation of $\hat{\Phi}_{xy}(\mu)$, the signals $x(k)$ and $y(k)$ are previously aligned to compensate for the coding delay. Similarly to [11], the distance measure

$$d = \frac{1}{N_{\max} - N_{\min} + 1} \sum_{\mu=N_{\min}}^{N_{\max}} \left(|\hat{H}(\mu)|^2 - 1 \right)^2 \tag{2}$$

is then defined with $N_{\min}$ and $N_{\max}$ corresponding to the frequency indices approximating the lower and the upper cut-off frequencies (50 Hz, 7000 Hz) of the codec, respectively. Since the estimated transfer function in (1) describes how much of the input $x(k)$ is linearly mapped to the output $y(k)$, the distance measure with an allpass as reference in (2) is interpreted as a measure for the degree of linearity/nonlinearity of the codec.

The expected trend of the transfer functions as defined in (1) is depicted in Fig. 3 for all available bitrates of the *AMR-WB* codec. For these plots, averaging was performed over the transfer functions computed from signal pairs based on the training corpus of [13]. The corresponding expected values of the distance measure (2) are given in Table I. From this

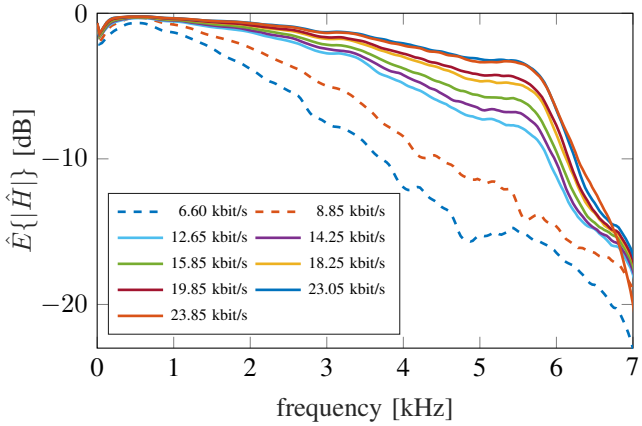

Fig. 3. Expected per-file transfer functions according to (1) of the *AMR-WB* codec for all available bitrates.

data it may be concluded that the proposed feature possesses discriminative properties regarding the detection of the bitrates of the *AMR-WB* codec.

III. Codebook Correlation

The distance measure proposed in the previous section depends on the availability of a reference signal $x(k)$. However, the feature is, therefore, able to incorporate information on the entire transmission system from beginning to end. This is of special interest if the speech is coded and decoded more than once in the transmission system. In contrast to that, this section focuses on a feature type which only requires a recorded signal $y(k)$ for its computation. Consequently, features of this type are specifically targeted at extracting information on the codec last applied by the transmission system. For the description of these features, the bitrate mode index i with $1 \leq i \leq 9$ is introduced. The bitrate mode index is directly proportional to the bitrate of the *AMR-WB* codec where $i = 1$ refers to the lowest and $i = 9$ to the highest available bitrate.

The *AMR-WB* codec is based on the principles of *Algebraic Code-Excited Linear Prediction (ACELP)* [14]. In accordance with [8], each subframe $\mathbf{u}(\lambda) = [u_1(\lambda), ..., u_{64}(\lambda)]^{\mathrm{T}}$ of the lower band excitation signal (50 - 6400 Hz) used for linear predictive synthesis is, therefore, the weighted sum of an adaptive codebook vector $\mathbf{v}(\lambda)$ and a fixed codebook vector $\mathbf{c}(\lambda)$:

$$\mathbf{u}(\lambda) = \hat{g}_p \mathbf{v}(\lambda) + \hat{g}_c \mathbf{c}(\lambda). \tag{3}$$

Since the fixed codebook vector is one of M_i predefined sequences $\mathbf{c}_j^{(i)}$ unique to the bitrate mode i with $1 \leq j \leq M_i$, the bitrate may be detected by identifying the sequences used for speech synthesis. A similar approach was used in [7] to distinguish between various *ACELP* codecs, but not their bitrates. For such an approach it is required to assume $\hat{g}_p \ll \hat{g}_c$, which is approximately true for unvoiced speech, to yield

$$\mathbf{u}(\lambda) \approx \hat{g}_c \mathbf{c}(\lambda) \; \forall \lambda \in \text{unvoiced}. \tag{4}$$

Hence, the preprocessing of $y(k)$ with an algorithm for the detection of unvoiced speech frames followed by linear pre-

ISBN 978-90-827970-1-5

TABLE I
AVERAGE PER-FILE FEATURE VALUES AT ALL CODEC BITRATES

kbit/s	6.60	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
$\hat{E}\{d\}$	0.52	0.41	0.26	0.23	0.20	0.15	0.14	0.09	0.10
$\hat{E}\{\varphi^{(1)}\}$	0.93	0.79	0.79	0.79	0.78	0.79	0.79	0.79	0.79

dictive analysis is assumed for the computation of the estimate $\hat{\mathbf{u}}(\lambda)$.

For sequence identfication, the correlation coefficient

$$\varphi_j^{(i)}(\lambda) = \frac{\hat{\mathbf{u}}^{\mathrm{T}}(\lambda) \cdot \mathbf{c}_j^{(i)}}{\|\hat{\mathbf{u}}(\lambda)\|_2 \cdot \|\mathbf{c}_j^{(i)}\|_2} \tag{5}$$

is utilized. It expresses the correlation of the excitation signal in subframe λ with the fixed codebook vector j of the codebook matched to the bitrate mode i. The coefficient

$$\varphi^{(i)}(\lambda) = \max_j \varphi_j^{(i)}(\lambda) \tag{6}$$

then expresses the maximum correlation of a subframe with a codebook. It was found that within a speech sample of about 5 s, the values

$$\varphi^{(i)} = \max_{\lambda \in \text{unvoiced}} \varphi^{(i)}(\lambda) \tag{7}$$

provide discriminative properties.

Fig. 4 depicts distributions of the feature value defined in (7) with regard to the codebook of the lowest bitrate. Two separate distributions are given, one for speech signals actually coded at the lowest bitrate and one for speech signals coded at any other bitrate. The speech was taken from the training corpus of [13]. For completeness, the mean values for the individual bitrates are given in Table I. The discriminative properties of (7) can be clearly observed.

Especially for the higher bitrates, the number M_i of available codebook vectors is significantly large. Consequently, the complexity for computing (6) needs to be reduced to allow for the desired application. By exploiting the codebook structure, the complexity of the computation may be rendered independent of M_i.

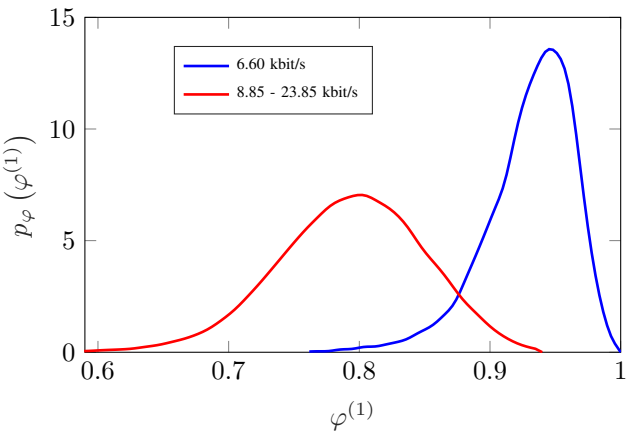

Fig. 4. Estimated per-file feature value distribution for $\varphi^{(1)}$.

All codebooks of the *AMR-WB* codec have in common that a bitrate-specific number of signed pulses of magnitude 1 are distributed over the 64 subframe sample positions [8]. All other positions are set to zero. The potential positions of the pulses are divided into tracks; two tracks $\mathbf{t}_\mathrm{e}$ and $\mathbf{t}_\mathrm{o}$ corresponding to even and odd sample indices, respectively, for the lowest bitrate and four tracks $\mathbf{t}_\kappa = [\kappa, 4+\kappa, ..., 60+\kappa]$ with $1 \leq \kappa \leq 4$ containing every fourth position for all other bitrates.

At the lowest bitrate, each track only contains one pulse and, thus, (6) may be rewritten as

$$\varphi^{(1)}(\lambda) = \frac{\max_{\alpha \in \mathbf{t}_\mathrm{o}}\{|\hat{u}_\alpha(\lambda)|\} + \max_{\alpha \in \mathbf{t}_\mathrm{e}}\{|\hat{u}_\alpha(\lambda)|\}}{\|\hat{\mathbf{u}}(\lambda)\|_2 \cdot \sqrt{2}} \tag{8}$$

due to the sifting properties of the vector multiplication in (5). Since $M_1 = 4096$, the computation of the maximum correlation by means of (8) yields a reduction of computational complexity of approximately 97 %.

For all other bitrates, the four tracks contain a both bitrate and track-specific number of pulses $n_\kappa^{(i)}$. For each track the vector $\mathbf{s}_{\text{desc},\kappa}(\lambda) = [s_{\kappa,1}(\lambda), ..., s_{\kappa,16}(\lambda)]$, which contains the absolute values of all samples of a track sorted in descending order, is defined. With this definition, the correlation coefficient may then be computed as

$$\varphi^{(i)}(\lambda) = \frac{\sum_{\kappa=1}^{4} \sum_{\alpha=1}^{n_\kappa^{(i)}} s_{\kappa,\alpha}(\lambda)}{\|\hat{\mathbf{u}}(\lambda)\|_2 \cdot \sqrt{\sum_{\kappa=1}^{4} n_\kappa^{(i)}}} \tag{9}$$

for $2 \leq i \leq 9$. The complexity of (9) is, obviously, not proportional to the total number M_i of available codevectors as is the case in (6). Thus, the computational complexity is reduced by at least 99.99 % for all applicable bitrates. When computing (7) by means of (9), complexity may be, additionally, reduced by writing the bitrate-specific normalization term outside of the maximum operation:

$$\varphi^{(i)} = \frac{1}{\sqrt{\sum_{\kappa=1}^{4} n_\kappa^{(i)}}} \max_{\lambda \in \text{unvoiced}} \frac{\sum_{\kappa=1}^{4} \sum_{\alpha=1}^{n_\kappa^{(i)}} s_{\kappa,\alpha}(\lambda)}{\|\hat{\mathbf{u}}(\lambda)\|_2}. \tag{10}$$

The correlation coefficient for the lowest bitrate may be formulated analogously using (8). The mathematical equivalence of (7) and (10) was simulatively confirmed.

IV. BITRATE CLASSIFICATION

The previous sections introduced features based on codec linearity/nonlinearity and on codebook correlation. This section now applies said features for a bitrate detection by means of a Gaussian-Mixture-Model (GMM) classifier. The choice of this classifier was primarily motivated by the Gaussian-like

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE II
RECOGNITION RATES FOR THE BITRATE CLASSES UNDER DIFFERENT CONSIDERATIONS OF FRAME OFFSET

(a) No Frame Offset

		Estimated class		
		B_1	B_2	B_3
True class	B_1	99.11 %	0.77 %	0.12 %
	B_2	0.60 %	98.15 %	1.25 %
	B_3	0.00 %	1.30 %	98.70 %

(b) Frame Offset in Testing Only

		Estimated class		
		B_1	B_2	B_3
True class	B_1	97.98 %	1.95 %	0.07 %
	B_2	13.75 %	82.32 %	3.94 %
	B_3	0.12 %	2.88 %	97.00 %

(c) Frame Offset in Training and Testing

		Estimated class		
		B_1	B_2	B_3
True class	B_1	95.44 %	4.52 %	0.04 %
	B_2	5.11 %	92.05 %	2.84 %
	B_3	0.01 %	3.01 %	96.98 %

feature value distributions (Fig. 4) observed for most of the features. This does, of course, not rule out that other classifiers may achieve comparable results.

While the *AMR-WB* codec does in fact provide nine different operational modes and, therefore, nine different bitrates, such a fine class distinction cannot be motivated from a speech quality point of view. As was shown in [9], the seven highest bitrates all offer a comparably high speech quality, whereas the two lowest bitrates yield a distinguishable quality degradation. Consequently, this work aims at distinguishing the two lowest bitrates (class B_1 with 6.60 kbit/s and class B_2 with 8.85 kbit/s) from the seven remaining bitrates (class B_3 with 12.65 - 23.85 kbit/s). A representative set of example speech files from the different classes is provided online by the authors [15].

Due to above class definition, the correlation features regarding the codebooks of bitrate modes 1-3 are used. The combination of these features with the linearity/nonlinearity measure then yields three 4-dimensional GMMs. The models were trained and tested with the training (4620 reference files) and testing (1680 reference files) portion of the TIMIT database [13], respectively. Each file consists of one spoken sentence. The coding and decoding of the signals was implemented using [16].

As listed in Table II, the proposed classifier achieves an identification rate of 98.7 % for the three defined bitrate classes. For these results, knowledge of the subframe positions in both training and testing was assumed for the computation of (7). With randomly offset testing sequences, the identification rate drops to 92.4 %. However, this effect may be more or less compensated by also allowing a random frame offset for the training sequences, yielding a 94.8 % identification rate.

If only the correlation features, which do not require the reference signal, are used in the statistical models, an overall identification rate of roughly 80 % may still be achieved. However, the application of these models yields a significant false positive rate regarding bitrate classes B_1 and B_2. Consequently, only the combination of both feature types provides the high classification accuracy required for the desired application.

V. SELF-TANDEM CLASSIFICATION

The occurrence of self-tandeming or transcoding yields a notable speech quality degradation [9]. Hence, it is desirable to identify these scenarios as the cause for an observed quality degradation. By modifying the proposed bitrate classifier, certain self-tandem types may be distinguished. For this, the classes

$$T_1 = \{B_3, B_3 \times B_3\}$$
$$T_{23} = \{T_2, T_3\}$$

with subsets

$$T_2 = \{B_1, B_2, B_1 \times B_3, B_2 \times B_3\}$$
$$T_3 = \{B_1 \times B_1, B_2 \times B_2, B_1 \times B_2\}$$

are defined. The crossproduct of two bitrate classes is to signify a self-tandem scenario with the applicable bitrates where the order of application is irrelevant. Class T_1 describes conditions with satisfactory speech quality since only high bitrates (B_3) are considered. This is in contrast to T_{23}, where at least one low bitrate is involved in the processing of each signal. The worst quality is achieved by signals in subset T_3, since all signals are subject to two consecutive low-bitrate coding and decoding processes.

By training a new set of statistical models for these classes analogously to the previous section, a recognition rate of 94.6 % is achieved regarding the distinction of T_1 and T_{23}. Consequently, the coding scenarios yielding a satisfactory speech quality are separated from the remaining scenarios with a high accuracy. Given the recognition of class T_{23}, the subsets T_2 and T_3 are then identified correctly with a rate of 82.1 %, providing additional information on the cause of suboptimal speech quality. In the context of network testing, where several sample recordings are available and no per-file results are required, this classifier, thus, provides a means to identify certain network scenarios regarding bitrate or self-tandem.

VI. CONCLUSION

This paper provided two different types of features for the classification of speech regarding coding effects within the application of network testing. It was shown that with these features a highly accurate bitrate detection for the *AMR-WB* codec is possible. This is done with only four signal features in total. Furthermore, since the features analyze the transmission system on different scales, it is also possible to gain information about the occurrence of self-tandem in the transmission system.

ISBN 978-90-827970-1-5

REFERENCES

[1] *Perceptual evaluation of speech quality (PESQ)*, ITU-T Rec. P.862.

[2] *Perceptual objective listening quality assessment (POLQA)*, ITU-T Rec. P.863.

[3] Deutsche Telekom, "Draft requirement specification for P.AMD (Perceptual Approaches for Multi-Dimensional Analysis)," ITU-T, 2011.

[4] F. Köster, F. Schiffner, S. Möller, and L. Malfait, "Towards degradation decomposition for voice communication system assessment," *Quality and User Experience*, 2017.

[5] F. Jenner and A. Kwasinski, "Highly accurate non-intrusive speech forensics for codec identifications from observed decoded signals," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 1737–1740.

[6] D. Sharma, P. A. Naylor, N. D. Gaubitch, and M. Brookes, "Non intrusive codec identification algorithm," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4477–4480.

[7] J. Zhou, D. Garcia-Romero, and C. Espy-Wilson, "Automatic speech codec identification with applications to tampering detection of speech recordings," in *Proc. Interspeech*, 2011, pp. 2533–2536.

[8] *Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*, ITU-T Rec. G.722.2.

[9] "Performance characterization of the Adaptive Multi-Rate Wide-Band (AMR-WB) speech codec," 3rd Generation Partnership Project, TR 26.976, 2002.

[10] U. Halka, "Objektive Qualitätsbeurteilung von Sprachkodierverfahren unter Anwendung von Sprachmodellprozessen," Ph.D. dissertation, Bochum University, 1993.

[11] T. Ludwig, "Messung von Signaleigenschaften zur referenzfreien Qualitätsbewertung von Telefonbandsprache," Ph.D. dissertation, Kiel University, 2003.

[12] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, Jun 1967.

[13] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," Philadelphia Linguistic Data Consortium, Tech. Rep. LDC93S1, 1993.

[14] C. Lamblin, "Algebraic code-excited linear prediction speech coding method," U.S. Patent 5,717,825, 1998.

[15] "Additional material," 2018. [Online]. Available: https://dss.tf.uni-kiel.de/index.php/research/publications/publications-add-material/bitrate-detection-arm-wb-codec

[16] "ANSI-C code for the Adaptive Mutli-Rate - Wideband (AMR-WB) speech codec," 3rd Gerneration Partnership Program, TS 26.173, 2001.

ISBN 978-90-827970-1-5