Assessment of Prosodic Attributes in Codec-Compressed Speech

Johanna Dobbriner, Oliver Jokisch, Michael Maruschke

Hochschule für Telekommunikation Leipzig, Institut für Kommunikationstechnik, Germany

{johanna.dobbriner,jokisch,maruschke}@hft-leipzig.de

Abstract

This article deals with the representation of prosodic attributes in coded speech which is less-studied. Common models in speech coding assume that there is no relevant influence of prosodic variation on perceived quality and content of coded speech under suitable operating conditions. Our experiments included a listening test and the instrumental assessment of utterances from an especially constructed test database for the three categories *focus*, *type of sentence* and *situation*. Each category contained at least three different text phrases in several variants, and each original sample was compressed using the fullband-audio Opus codec and the narrowband G.711 codec for reference. The listeners evaluated the overall speech quality and processed a matching task to given prosodic categories. In general, the prosodic variations were well-recognised even when the coding degradation was significant. The overall assessments were comparably high, by achieving an MOS of 4.3 and above on the five-point scale. The hybrid Opus coding method seems to maintain the prosodic features of speech as given in the original reference.

Index Terms: speech coding, prosody, listening test, MOS, POLQA, Opus, G.711

1. Introduction

Audio encoding and decoding is frequently used for the efficient transmission and storage of speech or music data and may influence the perceptual audio or speech quality. It is common to analyse quality from several perspectives, e. g. with regard to certain specific features of speech. This contribution addresses the prosodic attributes of coded speech which are less-studied so far.

We started with a more general assumption from a former study [1] that there is no significant influence of prosodic variation on the quality of adequately coded speech (i. e. using a proper bandwidth/bit rate).

Our experiments incorporated listening tests and a quality assessment using the instrumental POLQA method [2]. The speech data in the prosodic part of these experiments [3] consist of multiple utterances of four speakers (two male and two female) for the three categories *focus*, *type of sentence* and *situation*. Each category contained at least three different phrases in two or more variants. Each original sample was compressed using the Internet-based Opus codec [4] as well as the G.711 a-law codec [5] – two widely used representatives of different coding algorithms, which are generally known to compress speech at high quality in their respective domains.

Our listeners were asked to evaluate the overall speech quality, to correctly match it to a given variant and finally to assess their own assessment difficulty. This test was presented online via the Web platform Percy [6] to 14 participants.

2. Speech Coding Methods

There are a multitude of codecs for speech and audio compression, and new algorithms are continually developed. In general, it is possible to categorise these codecs by the frequency band with four commonly used bandwidths for speech and audio signals. Narrowband (NB) includes speech signals up to 3.4 kHz and is common in landline and mobile telephony. In wideband (WB), audio signals up to 7 kHz are used which is known as high definition (HD) voice. Super-wideband (SWB) comprises an extension of wideband, including signals that contain frequencies up to 14 kHz. Beyond, there is fullband (FB), wherein the full audible frequency range up to 20 kHz is included. This frequency bandwidth is commonly used on CD and known as full HD voice. For each bandwidth there are many codecs available using vastly different methods of audio compression. Among the least complex algorithms is the pulse-code-modulation, a sample-based type of waveform coding wherein each audio signal is sampled at first, and every sample is then quantised and coded. This is a low-delay method with a constant and relatively high bitrate mostly used in landline communication. There are also algorithms which use linear prediction, transform coding or a combination of various coding methods to compress speech at a high quality with short delay and a low bitrate. Among the more frequently used codecs are the G.711 codec which sets a quality standard for NB coding and the Adaptive Multi Rate (AMR) wideband codec in the WB frequency domain (e. g. in modern smartphones).

More recently, the FB Opus codec has become popular used for Web browser-based Real Time Communication (WebRTC). It enables users to easily communicate in full HD voice quality via internet browsers (e. g. Google Chrome). Out of the available codecs we chose two typical ones for our experiments, mainly due to the reported high quality in their respective domains – the G.711 a-law codec and the Opus codec.

2.1. G.711 codec

One of the codecs that has been in use for a very long time, since the 1970s, is the G.711 codec [5]. It was standardised in 1988 by the International Telecommunication Union, Telecommunication Standardization Sector (ITU-T) and is customary for landline communication. The coding method for G.711 is PCM with two different quantisation schemes, the μ-law quantisation which is common for example in the US and the a-law used in European telephony. G.711 has proven to produce high quality speech for NB signals and has an extremely low delay that makes it especially practical for real-time communication, even in other application areas. Nevertheless, its high bitrate of 64 kbps is inconvenient.

2.2. Opus codec

In 2012, the Internet Engineering Taskforce (IETF) standardised the novel audio codec Opus in the RFC 6716 [4]. Opus was designed as a highly-adaptive codec usable for broad application scenarios from speech to music, like VoIP, video conferencing or online gaming. It can work in all four different frequency bands at varying bitrates from 510 kbps to 6 kbps – achieving high quality audio either mono or stereo. Even the coding delay remains relatively bearable – from 2.5 to 60 ms depending on the use case. This adaptability to virtually any scenario is achieved by using a combination of several existing coding algorithms like SILK based on linear predictive coding and the CELT which uses the Modified Discrete Cosine Transform to compress audio signals.

3. Speech Quality Measures

In order to evaluate how well a codec performs in terms of speech quality, it is necessary to perform several tests. Depending on the aim of this assessment, there are various methods to assess speech quality. Quality is generally very subjective – it is the human user who decides to either use or not to use an application, regardless of what a theoretical model may have predicted. Therefore we employed two kinds of speech quality measures. The more important one was the listening test with human participants who separately evaluated a number of speech samples. The second type consisted of instrumental measures. These are algorithms designed to simulate human perception and thereby to predict the quality assessment, a listening test would result in. The listening tests can be distinguished in category/numerical and intelligibility tests.

3.1. Category rating

In the category test, listeners will be exposed to the speech samples and then rate the perceived quality on a scale. There is Absolute Category Rating (ACR), wherein the proband listens to single audio samples and assesses the quality of the samples on a numerical scale like the five-point scale of the Mean Opinion Score which is frequently used – category 5 means excellent and 1 represents very poor speech quality. Afterwards the mean of the scores is determined for each codec and signifies the overall speech quality of that codec. Furthermore, there is Degradation Category Rating (DCR) in which each coded sample is directly compared to the original one, where the listeners assess how much the coded sample is degraded on a numerical scale. A third method is the Comparison Category Rating (CCR) which works similar to the DCR test but the two samples for comparison can be any of them – from two different codecs or codec and original – and the first sample heard is the reference, whereas the second must be rated as better, equal or worse to the reference on a numerical scale.

3.2. Intelligibility test

Intelligibility tests, on the other hand, are necessary in order to determine how clearly coded speech can be understood. The participant listens to a word or phrase and is asked to write down what he/she understood. Afterwards, the percentage of correctly understood speech samples is determined and interpreted as a measure of the intelligibility of speech coded with the tested algorithm. Typical representatives of this method are the Diagnostic Rhyme Test (DRT) and the Diagnostic Alliteration Test (DALT). In both tests, several pairs of words are given and the listeners have to decide which word from the current pair they heard. The DRT uses words with similar endings like “milk” and “silk” whereas the pairs in a DALT begin similarly e. g. “arm” and “art”. In this kind of test only trained listeners take part usually.

3.3. Instrumental assessment

As listening tests are usually time consuming and require much effort and organisation, many developers use instrumental quality measures to evaluate e. g. smaller developments in the coding algorithm or certain aspects and scenarios of speech coding. These instrumental methods use algorithms and perceptual models to approximate the likely results instead of the according listening test. The methods are more cost and time efficient, but also less accurate than real listening tests.

An established method for an instrumental speech quality assessment is POLQA, the Perceptual Objective Listening Quality Assessment defined in P.863 by ITU-T in 2007 [2]. It was designed as an improvement of its predecessor method P.862, also known as PESQ, the Perceptual Evaluation of Speech Quality. The POLQA algorithm requires all speech samples to fit into certain conditions, e. g. regarding the sampling rate. There are two different operating modes to assess NB and SWB signals – both resulting in a MOS-like quality measure.

4. Experiments

To conduct our experiments, it was necessary to focus on certain parameters in terms of the codec selection and the choice of prosodic attributes for the analysis. Furthermore, we needed to generate a number of speech samples representing these attributes and to decide on a certain test design.

4.1. Codec selection

We selected the Opus codec [4] as an example of frame-based hybrid coding, because it is an up-to-date standard published by the IETF, and high-quality audio signals can be encoded and transmitted at comparatively low bit rates. In the case of this study, we decided to encode our original speech samples using the default settings of the Opus codec in WebRTC: a bit rate of 32 kbps and a sampling rate of 48 kHz in FB audio.

As second codec we selected the G.711 a-law codec [5], standardised by ITU-T, and a longstanding representative of sampling-based waveform coding in NB speech. It is often taken as a reference for speech quality in this bandwidth and has been in use since the 1970s.

4.2. Prosodic parameters

As it was our aim to focus on prosody, we chose three specific attributes of speech that can be expressed almost exclusively by prosody in German. Thus, it was possible to use the same wording combined with varying prosodic attributes to express different meanings, e. g. “Es regnet.” *(It’s raining.)* as opposed to “Es regnet?” *(It’s raining?)*. The three selected attributes were *focus*, *type of sentence* and *situation*, representing different categories of the experiment.

Focus contained phrases, where a different word was emphasised in each variation, whereas the category *type of*

sentence consisted of utterances that could be either question or statement, depending on intonation.

In *situation*, there were sentences which could be divided into different phrases which changed the situation expressed in these sentences. One example for this category is

"Max dachte, Lisa kommt aus Hamburg." *(Max thought, "Lisa is from Hamburg.")* or

"Max, dachte Lisa, kommt aus Hamburg." *("Max", thought Lisa "is from Hamburg.")*.

4.3. Speech data

We collected speech samples for each category and recorded four speakers (2 males aged 14 and 18, two females, 23 and 46) who uttered the variations and repetitions. In the end, we were able to choose from 220 utterances.

For the tests, we selected three sentences in two variations each for the category *type of sentence*. Four phrases in two or three variants were chosen for *focus* and another four sentences with two variations per phrase were used in the category *situation*.

4.4. Test design

The experiments incorporated both, listening test and quality assessment using the instrumental POLQA method [2]. While the POLQA testing required nothing else than representative samples, the listening test needed to be efficiently designed for human participants. Consequently, we used the Web-based platform Percy [6], an adequate form of distribution for this experiment.

Our listeners were asked to first evaluate the overall speech quality of the current sample, then to match it to one out of several given variants and finally to assess their own difficulty of choosing one variant. Within the categories, original and coded samples were presented in random order.

For the participants, the only difference between the categories was their choice of variants to match. In *type of sentence*, they had to choose between *question* and *statement*, in *focus* there were as many choices as there were words in each utterance, for example:

o Alle o Jungen o spielen o Fußball.

In *situation*, we drew two images per sentence, visualising both possible variants of one phrase with short descriptions below them, from which the listeners had to choose the one they heard, e. g.: "Das Schiff verließ den Hafen nicht ohne Benachrichtigung des Kapitäns."*(The ship didn't leave the harbour without notification of the captain.)*. In this example, either the ship did *not* leave and the captain wasn't notified (as shown in Figure 1) or the ship *did* leave, but not until the captain knew about it, which can be seen in Figure 2.

4.5. Listeners

The online listening test was distributed among native German listeners. In total, we had 20 participants, although only 14 of them actually completed the test – including eight male and six female listeners between 21 and 76 years (mean age of 34) from various regions of Germany.

The test environment was widely quiet (home or office) and involved different audio equipments. The average testing time resulted to ca. 30 min.

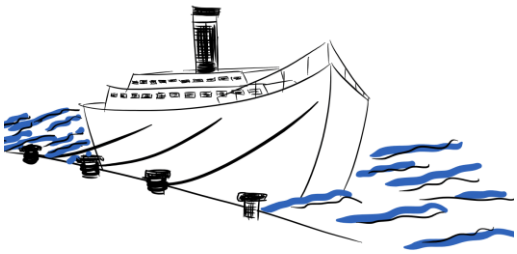

Figure 1: *Das Schiff verließ den Hafen NICHT.*

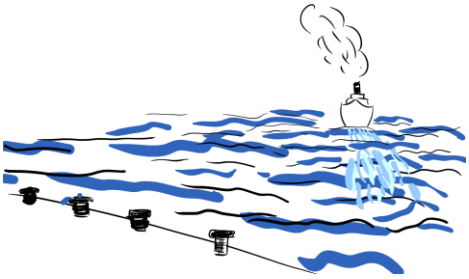

Figure 2: *Das Schiff VERLIESS den Hafen.*

5. Results

In the POLQA test using SWB mode, the Opus-coded samples achieved a score of 4.58 whereas for G.711 as a NB codec, this mode was not suitable.

The results of the listening test are shown in the Figures 3 and 4, first for the overall assessment and the second for the success rate of matching speech samples to given variants.

5.1. Overall speech quality

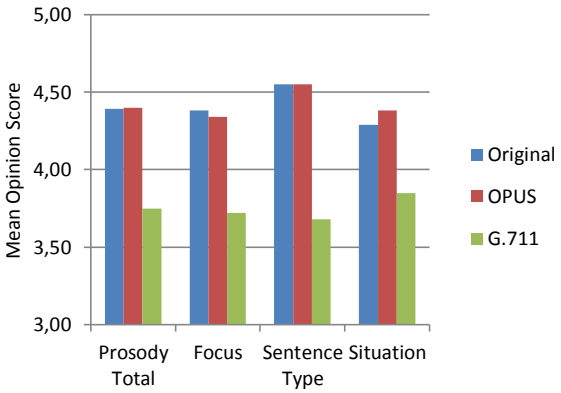

Figure 3: *Speech quality per category and codec*

Figure 3 shows the results of the MOS test when participants were asked to assess the speech quality of what they heard on the scale from 1 to 5. As expected, taking into consideration the different bandwidths of the speech signals, the results for the NB signals of G.711 are noticeably lower than those for the Opus coded FB samples and the original reference.

In the overall speech quality on the MOS scale, G.711 achieved a score of 3.75 whereas the Opus coded and original samples rated at 4.40 respectively 4.39. Interestingly,

averaged over all prosodic phrases, the Opus coding results seem marginally better than the original samples without any degradation.

Furthermore, the quality scores by prosodic category are shown. The categories, as described in the experimental setup are *focus*, *sentence type* and *situation*. In general, the speech samples in *sentence type* received the highest scores of 4.55 for both original and Opus-coded speech, whereas this category scored the lowest at 3.68 when coded with the G.711 codec.

For Opus codec, the other categories show no significant differences with 4.34 in *focus* and 4.38 in *situation*. G.711 performed different, where the *focus* samples were rated only slightly better than those from *sentence type* with a MOS of 3.72. The category *situation* on the other hand, received a rating of 3.85 which is significantly higher.

The original samples were rated higher than or equal to the Opus codec in the two categories *focus*, receiving 4.38, and *sentence type* with 4.55. The lowest rating for original samples is found in the *situation* category with an MOS of 4.29 only.

5.2. Category matching task

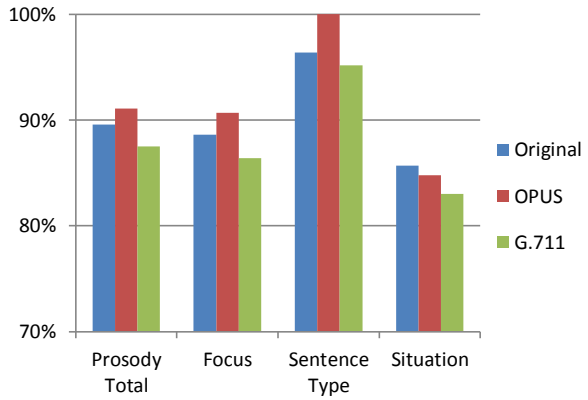

Figure 4: *Success rates within the categories*

In Figure 4, the success rates of the intelligibility test are presented. The *success rate* represents the percentage of samples that were matched to the correct variant depending on the category. These were the emphasised word in *focus*, either question or answer for *sentence type* and one image versus the other in case of *situation*.

In all coding methods, *sentence type* yielded the highest proportion of correct matches while differentiating between two *situations* resulted in the highest number of mistakes. Synchronously with the MOS results, G.711 samples were most frequently mismatched in comparison to the other two, but the relative gap between different coding methods is smaller than in the speech quality assessment. Overall, Opus scored 91.1 %, followed by the original samples at 89.6 % and G.711 succeeding in 87.5 % of the samples.

Overall, the success rates for prosody-only variations were rather high. Again, Opus outperformed the original samples in two of the three categories, namely *focus* and *sentence type* by 2.1 % and 3.6 % respectively and even in the *situation* category there was only a small difference of 0.9 % between both.

5.3. Matching effort

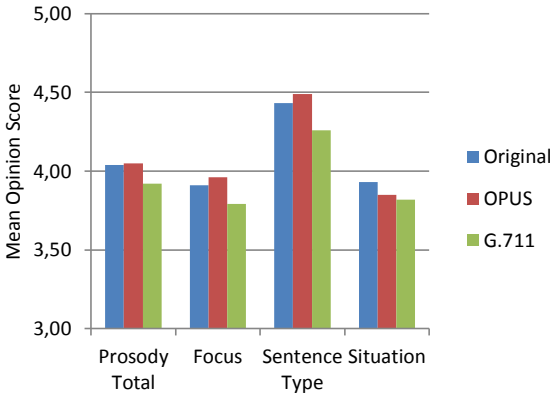

Figure 5: *Matching difficulty per category and codec*

Figure 5 displays the average difficulty of the listeners to match their samples to one of the given answers, using scale from 1 to 5, wherein 1 represents high effort/difficulty while 5 means that the listeners found the matching task easy. For a better comparison (and as in Figure 3 and 4), the results are listed by category and coding method.

The subjective *difficulty* of matching the speech sample to one variant showed a higher variability, although one constant was the category *sentence type* being again assessed with the highest scores of 4.49 for Opus, 4.43 for the original samples and 4.26 in case of G.711 coding (this time meaning that people typically found it easy to decide whether they had just heard a question or a statement). Additionally, this diagram shows once again that Opus and the originals received almost identical assessments of overall 4.05 and 4.04 respectively. G.711 scored worst at 3.92 but in this case with a smaller difference than in MOS or success rate evaluation. Except for Opus, the difficulty of matching for *focus* and *situation* was evaluated almost equal within each coding method: For Opus the *focus* category achieved a difficulty of 3.91 whereas in *situation* it was 3.93 and G.711 was assessed with 3.79 in *focus* and 3.82 in *situation*. Opus on the other hand, scored 3.96 in the category *focus* and 3.85 in *situation*.

6. Discussion and Conclusions

Comparing the MOS results for speech quality to the success rate, it is evident that the success rate scores are relatively higher than the scores in speech quality, which is especially significant in case of the G.711 codec. Therefore it would be one conclusion that prosodic variations are generally well recognised even when the overall speech quality is degraded significantly.

Out of the three prosodic categories, *type of sentence* proved to be the most easily one and also gained the highest scores independent of coding algorithm and setting. In general, there were only small assessment differences between the original samples and its coded equivalents, and in most cases the assessment of Opus-coded speech was even slightly higher compared to the assessment of original samples.

Overall, the assessments were comparably high – approx. 4.3 and above on the five-point MOS scale.

As a preliminary conclusion, the Opus algorithm seems to represent the prosodic features of speech as well as (or even slightly better?) than original speech data do, but further tests need to prove the significance of this finding.

Beyond, prosodic differences in questions and declarative sentences are recognized easier by listeners compared to other prosodic attributes of speech, regardless of external interferences. Our further research will focus on further prosodic and paralinguistic features, alternative coding algorithms and a larger set of validation data.

7. Acknowledgements

We would like to thank Christoph Draxler from LMU Munich for the opportunity to perform our experiment on the Percy Web platform and to our volunteer listeners. Further thank goes to the SwissQual AG, Zuchwil (a Rhode & Schwarz company in Switzerland) for supplying the POLQA testing software – in particular to Jens Berger.

8. References

[1] JOKISCH, O.; MARUSCHKE, M.; MESZAROS, M; IAROSHENKO, V.: Audio and speech quality survey of the opus codec in web real-time communication, Elektronische Sprachsignalverarbeitung. Tagungsband der 27. Konferenz (O. Jokisch, ed.), vol. 81 of Studientexte zur Sprachkommunikation, Leipzig, Germany, pp. 254–262, TUDpress, 2016.

[2] ITU-T: Methods for objective and subjective assessment of speech quality (POLQA): Perceptual Objective Listening Quality Assessment, REC P.863, International Telecommunication Union (Telecommunication Standardization Sector), Sept. 2014.

[3] DOBBRINER, J.: Beeinflussung prosodischer Sprachmerkmale durch Sprach- und Audiocodecs. Bachelorarbeit, Hochschule für Telekommunikation Leipzig, Mai 2016.

[4] VALIN, J; VOS, K.; TERRIBERRY, T.: Definition of the opus audio codec, RFC 6716 (Proposed Standard), Internet Engineering Task Force, Sep. 2012. [Online]. Available: http://www.ietf.org/rfc/rfc6716.txt.

[5] ITU-T: Pulse code modulation (PCM) of voice frequencies, REC G.711, International Telecommunication Union (Telecommunication Standardization Sector), November 1988. Available: https://www.itu.int/rec/T-REC-G.711

[6] DRAXLER, C.: Percy – An HTML5 Framework for Media Rich Web Experiments on Mobile Devices. Proc. 12th Interspeech Conference, pp. 3339 – 3340, Florence, August 2011.