A Model for the Perceptual Impact of Time-Clipping on the Discontinuity Dimension of Speech

Jovana Vranic[1], Christian Schmidmer[2]
[1] OPTICOM GmbH, Erlangen, Germany, E-Mail:JovanaV@opticom.de
[2] OPTICOM GmbH, Erlangen, Germany, E-Mail: cs@opticom.de

Abstract

Perceptual models for predicting the integral quality of transmitted speech are a reliable way to quantify the quality of audio devices or communication networks. Even though the overall speech quality represents an easily understood measure, it does not provide adequate information about the causes of the speech quality degradations. For that reason, models which predict individual dimensions of speech quality have been under development. The focus of this paper lies on the discontinuity dimension, which in previous studies has been shown to carry an important aspect in speech quality evaluation. We analyzed the influence of the *time-clipping rate* (N_{TC}) as well as the *total duration of time-clipping* (T_{tot}) on the perceived discontinuity to derive a perceptual time-clipping indicator (TC). Based on the subjective data, a linear relationship between the mean opinion score (MOS) for discontinuity dimension and the proposed TC indicator is observed. Furthermore, the proposed TC indicator was developed in the POLQA (Perceptual Objective Listening Quality Assessment) algorithm for the assessment of the overall perceptual quality. In this respect, the proposed algorithm yields high prediction accuracy for a variety of time-clipping conditions.

Introduction

State-of-the-art standardized instrumental methods, e.g., POLQA [4] provide highly accurate quality predictions for speech signals on 1 (Bad) to 5 (Excellent) MOS scale. Despite of its high accuracy for the overall quality, POLQA scores do not provide enough information about the causes of an insufficient speech quality. Note that such information in audio-quality-based troubleshooting is often quite important to correctly diagnose problems in the system (device, transmission network, etc.) under test.

Motivated by aforementioned applications of models which focus on individual dimensions rather than the overall quality the multidimensional nature of speech quality was investigated and four orthogonal dimensions of speech quality are defined in [1], namely *noisiness*, *discontinuity*, *coloration* and *loudness*. This paper focuses on the discontinuity dimension of the speech quality. In general, the discontinuity in transmitted speech signals can be further decomposed into three sub-dimensions [2]: *interruptedness*, *additive-artifacts* and *musical-noise*. This paper focuses on *interruptedness* sub-dimension and its relationship with the discontinuity dimension. In the following section, the discontinuity dimension and its sub-dimensions [2] are described in more details.

Based on the auditory tests [3], [5], [6] and the model proposed in [7], an improved time-clipping detection framework is developed. The developed algorithm uses the energies of samples in the reference and degraded speech signals to determine the time-clipping rate and total duration of time-clipping in an accurate way. Additionally, a time-clipping indicator (TC) which can be used to accurately estimate the discontinuity MOS is derived based on the analysis performed on training databases. The proposed time-clipping detection algorithm together with a TC is evaluated with POLQA for a broad set of time-clipping patterns using a dedicated validation database.

The Discontinuity Dimension

The discontinuity in the speech signal corresponds to either an isolated distortion or a non-stationary distortion [1]. Isolated distortions leading to *interruptedness* are caused either by the loss of packets during VoIP (voice-over internet protocol) transmissions or by erroneous bits incurred during radio transmission. Besides *interruptedness*, discontinuity could also be caused by imperfect noise reduction, which results in musical-noise, wherein, segment(s) of speech sound like a piece of music [2]. A third sub-dimension of discontinuity is additive-artifacts and is proportional to the frequency of the artifacts generated by, e.g., Packet Loss Concealment (PLC). As shown in Figure 1, an indicator for the discontinuity dimension should ideally take all of the above-mentioned distortions into account.

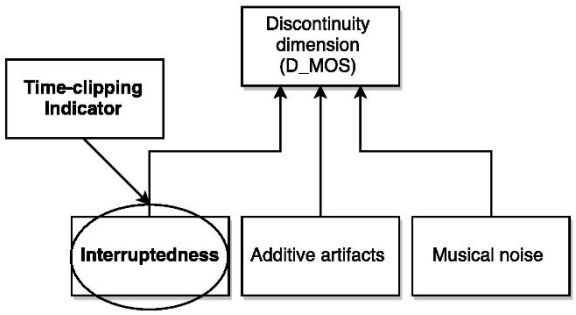

Figure 1 Sub-dimensions of discontinuity in speech quality.

Time-clipping Quality Indicator

Three databases have been used for the development and training of the TC indicator. Two of these databases are in German language (PAMD_A_DTAG3, PAMD_A_SQ1), whereas the 3rd database is in French language

(PAMD_A_Orange1). The subjective test procedure and requirements followed in case of the database PAMD_A_DTAG3 and PAMD_A_SQ1 are described in [6] and [3], respectively; while for PAMD_A_Orange1 they are described in [5]. These subjective tests were conducted in three different laboratories independently of this research. Based on the conditions containing time-clipping degradations in the three above-mentioned databases the TC indicator was formulated.

In each of the three training databases two time-clipping conditions were included, each represented by four speech samples. For the purpose of modeling the influence of time-clipping on the discontinuity dimension, we used these 24 speech samples together with their reference speech samples. The description of the conditions, which were used as the training set follows in Table 1.

Table 1 Training databases conditions

Condition	Description
1	Super-wideband
2	2% Time-clipping
3	20% Time-clipping

By listening to the degraded files and comparison with subjective results, it could be observed that two important characteristics of time-clipping degradations are influencing the discontinuity of transmitted speech. Both, time-clipping rate N_{TC} and total duration of time-clipping T_{tot} have to be taken into account when modeling *interruptedness* of transmitted speech. Additionally, the subjective data reveals that the discontinuity dimension subjective MOS (D_MOS) correlates quite well with the product of N_{TC} and T_{tot}. Since, a logarithmic dependency between the discontinuity and the product N_{TC}·T_{tot} has been observed, the final TC indicator was formulated as:

$$TC = 10 \cdot \log_{10}(T_{tot} \cdot N_{TC}) \qquad \text{[dB]} \quad (1)$$

Figure 2 plots the D_MOS from the 24 time-clipping and 12 reference samples and the time-clipping indicator (TC). For this plot the number of time-clipping sections (which defines N_{TC}) and total duration of time-clipping (which defines T_{tot}) were measured by visual inspection of the waveforms of the speech files and the databases were manually tagged with this information. Since the time-clipping was measured manually TC is referred to as $TC_{measured}$. As can be seen, the data indicate a linear dependency between D_MOS and the TC indicator. Based on this promising result, the aim became to develop an algorithm, which calculates the TC indicator directly based on speech signals. In the following section, an algorithm which computes the time-clipping specifics, namely N_{TC} and T_{tot} is detailed.

Time-clipping Detection Algorithm

The manual measuring of TC served as a first step to show that the proposed model correctly describes the influence of TC on discontinuity. In order to be able to detect and evaluate TC distortions in transmitted speech, the instrumental method has to be developed.

The developed time-clipping detection algorithm (Figure 3) compares the energy of the degraded signal (E_{deg}) with the energy of the reference signal (E_{ref}), for each frame (k), in order to check if the signal is time-clipped or not.

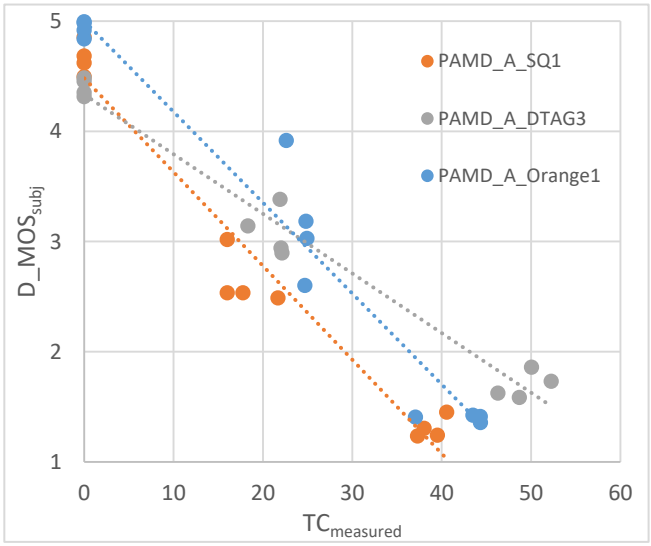

Figure 2 Linear relationship between the subjective discontinuity MOS with the manually measured TC indicator for the training databases.

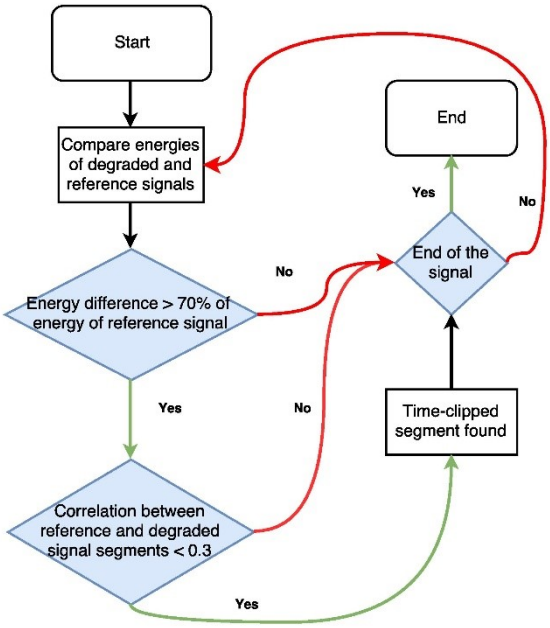

Figure 3 Flow chart of time-clipping detection algorithm

$$E_{ref}[k] - E_{deg}[k] > 0.7 \cdot E_{ref}[k] \qquad (2)$$

If the difference between the energies of the reference and the degraded signal for one frame is greater than 70% of the energy of the reference signal for the corresponding frame (Equation 2) and if the reference signal in that frame is audible, the counter of time-clipped frames is increased. If the number of the successive frames marked as time-clipped is greater than the given threshold (4 frames corresponding to 128 samples) and the Pearson correlation between the corresponding segments in reference and degraded signal is lower than 0.3, time-clipping is assumed. As Equation (1)

shows, the time-clipping indicator is finally calculated as a product of total duration of time-clipping in milliseconds (T_{tot}) and time-clipping rate (N_{TC}).

In the following, in Figure 4, the previously described time-clipping detection algorithm is evaluated for the three training databases, which were used for the model development. Figure 4 shows the subjective D_MOS of conditions detailed in Table 1 against the machine calculated TC indicator. Since the time-clipping was calculated using the detection algorithm it is referred to TC_{calc}. One can see that the subjective MOS values are almost ideally predicted by a linear function of the computed TC indicator.

Model Validation

In order to validate the behaviour of time-clipping detection algorithm and the derived indicator of Equation (1) for unknown test condition, a dedicated validation database (TC1) which focused only on discontinuity degradations was produced. This speech material contained files degraded by 10 different time-clipping patterns, as described in Table 2. Each time-clipping pattern was combined with four reference files (corresponding to two male and two female speakers) to produce 40 time-clipped speech signals. Reference files were chosen from the PAMD_A_DTAG3 database. Reference speech files are included as hidden reference in the subjective test. The details regarding the time-clipping simulated in the validation database is given in Table 2. Different time-clipping rates and different total durations of time-clipping have been chosen such that the expected TC indicator values are spread nearly uniformly on the scale from 0 to 40 dB.

Table 2 Conditions for the validation database (TC1)

Condition	Characteristics	
	N_{TC}	T_{tot} in milliseconds
1	1	5
2	1	10
3	1	50
4	2	50
5	4	50
6	1	400
7	2	400
8	4	400
9	8	400
10	16	400

A subjective listening test with 10 expert listeners was conducted in order to collect subjective scores for the discontinuity MOS. The subjects were asked to assess only the discontinuity and not the overall quality of the played speech. The files could not be repeated and were played in a random order for each test subject. Scoring was on a five point MOS scale, 1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent.

It is important to note that while the training databases contained a broad range of distorts apart from time-clipping, the validation database TC1 contained time-clipping only. Figure 5 shows the resulting subjective discontinuity MOS scores against the manually tagged TC indicator for the validation database. Auditory scores given in are averaged per condition for all four speech sources used for

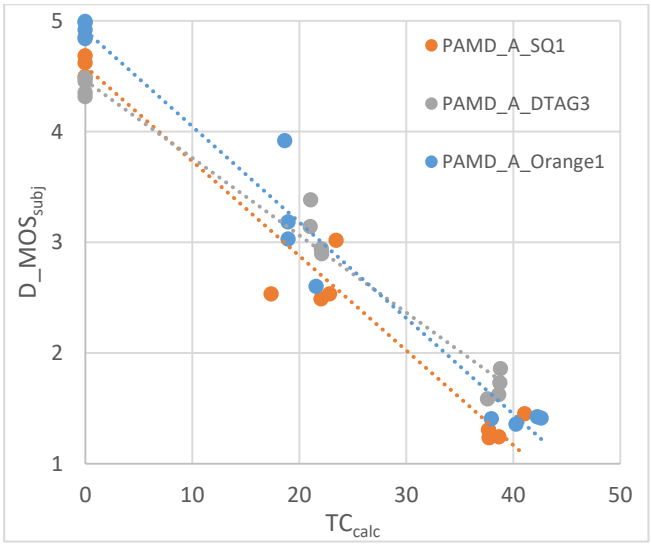

Figure 4 The subjective discontinuity MOS values against the program-based calculated TC indicator values per file for training databases.

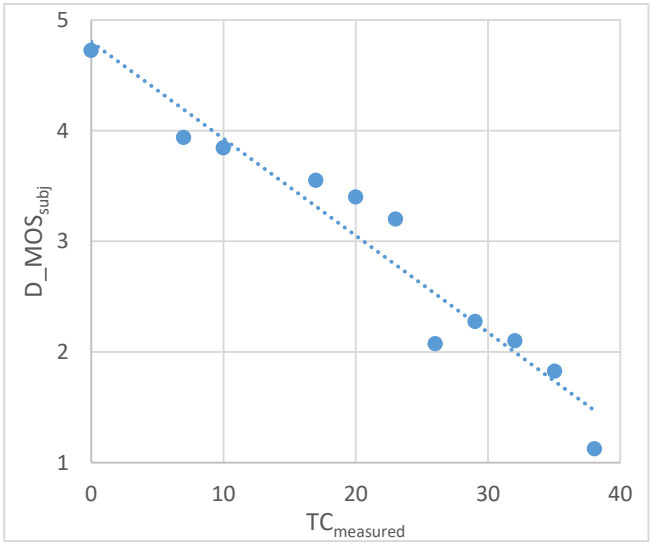

Figure 5 Linear relationship between the subjective discontinuity MOS and the product of time-clipping rate and overall duration of time-clipping given in dB, using manually measured clipping duration and clipping frequency information for the validation database TC1.

that condition. As the data in the training database indicated, the discontinuity MOS depends linearly on the TC indicator. This dependency has been validated for the TC1 database.

The automatic detection of time-clipping together with indicator of Equation (1) has been validated based on the speech samples of the newly created validation database in Figure 6. Similar to Figure 5 auditory scores given in are averaged per condition for all four sources used for that condition. The high correlation of the proposed indicator with the discontinuity MOS reveals the potential of the time-clipping detection algorithm together with the time-clipping indicator for modeling discontinuity in a pure objective sense, measured directly from the degraded speech signals.

Finally, Figure 7 shows the comparison between the manually measured and the machine calculated TC indicator for the validation database averaged per condition. As can be seen from the figure, the algorithm gives quite similar results compared to the manual measurement of time-clipping parameters. Thus, together with the TC indicator, it can be used to evaluate discontinuity for application scenarios wherein time-clipping can occurs as a part of the process in the transmission/processing chain.

Conclusions

In this paper, we proposed a time-clipping indicator as an accurate measure of *interruptedness* in transmitted speech. A linear relationship between the proposed indicator and the discontinuity dimension of perceptual speech quality has been observed and validated. Furthermore, an algorithm to directly measure the time-clipping indicator from real speech signals was developed.

During the work on this indicator, it was observed that distinguishing between voiced and unvoiced speech when estimating the influence of time-clipping on the discontinuity of the transmitted speech may further improve the accuracy of the prediction. It is therefore planned taking this into account in future research. Eventually, in order to have an estimator of the discontinuity dimension other sub-dimensions attributing to discontinuity, such as musical noise and additive artifacts, will be modeled as a part of the future work as well.

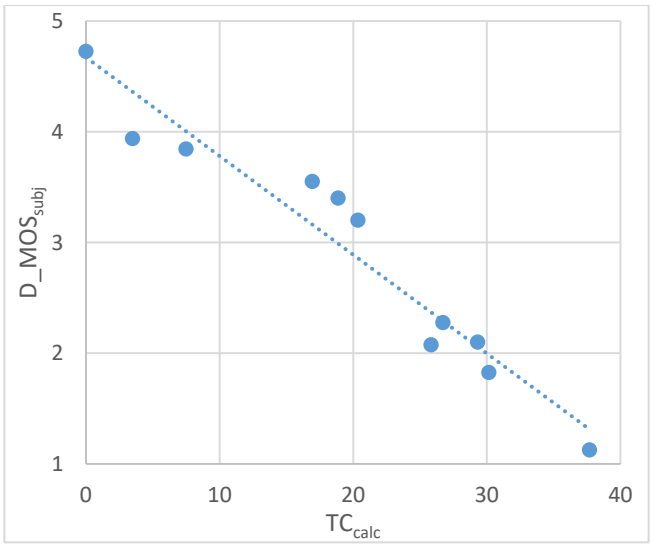

Figure 6 The subjective D_MOS against the estimated TC indicator averaged per condition for the validation database.

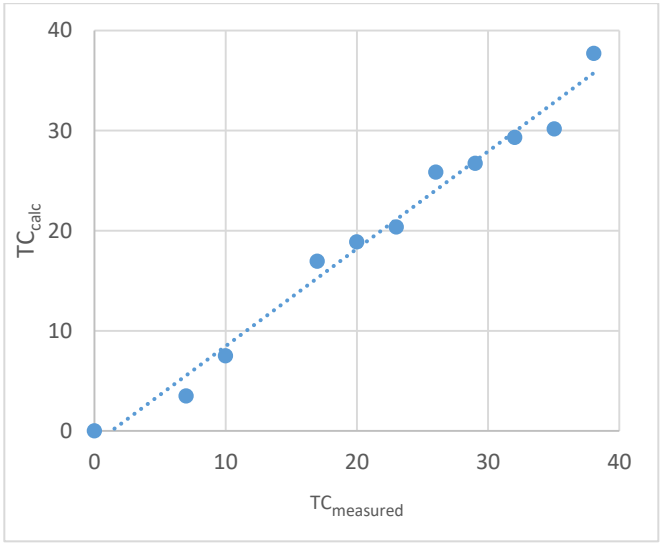

Figure 7 The linear relationship between machine-based estimated TC indicator and manually measured TC indicator averaged per condition including the reference speech samples for the validation database TC1.

References

[1] Cote N. *Integral and Diagnostic Intrusive Prediction of Speech Quality*, Springer, Berlin, 2011.

[2] Huo L, Wältermann M, Heute U, and Möller S. "Estimation of the speech quality dimension "Discontinuity"," *8th ITG Fachbericht-Sprachkommunikation*, pp.1-4, 2008.

[3] ITU-T Contribution COM12-C308-E, *P.AMD experiment for Set A*, International Telecommunication Union, Geneva, 2015.

[4] ITU-T Recommendation P.863, *Perceptual Objective Listening Quality Assessment*, International Telecommunication Union, Geneva, 2014.

[5] ITU-T Contribution COM12-C287-E, *Subjective test from Orange for the P.AMD project set A*, International Telecommunication Union, Geneva, 2015.

[6] ITU-T Contribution COM12-C195-E, *Draft requirements specified for the P.AMD (Perceptual approaches for Multi-dimensional Analysis)*, International Telecommunication Union, Geneva, 2011.

[7] Vranic J. "Validation and Development of Distortion Dimension Indicators for Noisiness, Loudness and Discontinuity for Super-Wideband Speech Signals", Master thesis, University Erlangen-Nürnberg, 2016.