Monitoring the Effects of Temporal Clipping on VoIP Speech Quality

Andrew Hines[1], *Jan Skoglund*[2], *Anil Kokaram*[2], *Naomi Harte*[1]

[1]Sigmedia, Trinity College Dublin, Ireland
[2]Google, Inc., Mountain View, CA, USA

andrew.hines@tcd.ie

Abstract

This paper presents work on a real-time temporal clipping monitoring tool for VoIP. Temporal clipping can occur as a result of voice activity detection (VAD) or echo cancellation where comfort noise in used in place of clipped speech segments. The algorithm presented will form part of a no-reference objective model for quantifying perceived speech quality in VoIP. The overall approach uses a modular design that will help pinpoint the reason for degradations in addition to quantifying their impact on speech quality. The new algorithm was tested for VAD compared over a range of thresholds and varied speech frame sizes. The results are compared to objective Mean Opinion Scores (MOS-LQO) from POLQA. The results show that the proposed algorithm can efficiently predict temporal clipping in speech and correlates well with the full reference quality predictions from POLQA. The model shows good potential for use in a real-time monitoring tool.

Index Terms: temporal clipping, VAD, VoIP, POLQA

1. Introduction

Speech communication using voice over internet protocol (VoIP) services such as Google Hangouts and Skype is becoming more pervasive and, along with mobile telephony, is gradually replacing traditional fixed line narrowband telephony. Monitoring the Quality of Experience (QoE) for users of these systems in realtime has become more complex as the points of failure have expanded.

Two commonly used speech enhancement algorithms are Voice Activity Detection (VAD) and line echo cancellation. VAD is implemented within most common speech codecs [1, 2] to suppress silence segments and to reduce unnecessary bandwidth consumption. The performance depends on the implementation but when a VAD algorithm misclassifies a portion of active speech as non-active, temporal clipping occurs and can impact the overall speech quality perceived by the listener. Echo cancellation is handled in a similar manner where segments containing reflected echo speech are suppressed. Both systems usually replace the inactive or echo segments with comfort noise [3].

Traditionally, QoE for voice communication systems is assessed in terms of speech quality. Subjective listener tests use an absolute category rating to establish a mean opinion score (MOS) on a five point scale by evaluating speech samples in laboratory conditions. Aside from being time consuming and expensive, these tests are not suitable for realtime monitoring of systems.

The development of objective models that seek to emulate listener tests and predict MOS scores is an active topic of research and has resulted in a number of industry standards. Models can be categorised by application, i.e. planning, optimisation, monitoring and maintenance [4]. Full reference objective models, such as PESQ [5] and POLQA [6], predict speech quality by comparing a reference speech signal to a received signal and quantifying the difference between them. Such models can be applied to system optimisation but are constrained by the requirement to have access to the original signal, which is not always practical for realtime monitoring systems. In these scenarios, no reference (NR) models, such as P.563 [7], LCQA [8] or ANIQUE+ [9] are more appropriate. They are sometimes referred to as single ended, or non-intrusive models, as they attempt to quantify the quality based only on evaluating the received speech signal without access to a clean reference. This restriction makes NR model design more difficult, and NR models tend to have inferior performance accuracy, when compared to full reference models [10]. Full reference metrics are sensitive to quality degradation caused by temporal clipping. A non-intrusive method of temporal clipping detection has been proposed by Ding et al. [11] but assumes prior knowledge of the clipping statistics. This new method attempts to monitor in a no-reference way that does not require access to the original unclipped signal or any knowledge of typical clipping statistics. Thus this is the first work we are aware of on full non-reference detection of clipping for speech quality measurement.

This work presents the early stage development of temporal clipping detection algorithm as one module of an overall NR speech quality model for VoIP application. The final model will contain multiple modules designed to detect and estimate the amount of degradation caused by specific VoIP degradations. Ultimately the output of individual modules will be combined to produce an aggregate objective speech quality prediction score. The novelty of this approach over other NR models [7, 8, 9] is that each module provides a unidimensional quality index feeding into the overall metric but can also provide diagnostic information about the cause of the degradation for narrowband or wideband speech. This will allow realtime remedial action to improve the overall quality of experience for the users of VoIP systems, through changing parameters during a call. For example, bandwidth may be adjusted to switch the quality of experience from a low-quality wideband speech scenario to a superior high-quality narrowband speech scenario.

The temporal clipping detection algorithm proposed in this paper, as part of the overall system, is designed to work with narrowband and wideband signals. While the tests presented here focus on narrowband signals, the algorithm could equally be applied to wideband scenarios. Section 2 describe temporal clipping and the causes. Section 3 describes the model and an experimental evaluation is outlined in section 4. Section 5 discusses the results and compares them with the predictions from POLQA. The paper concludes with a description of the next stages in the overall model development.

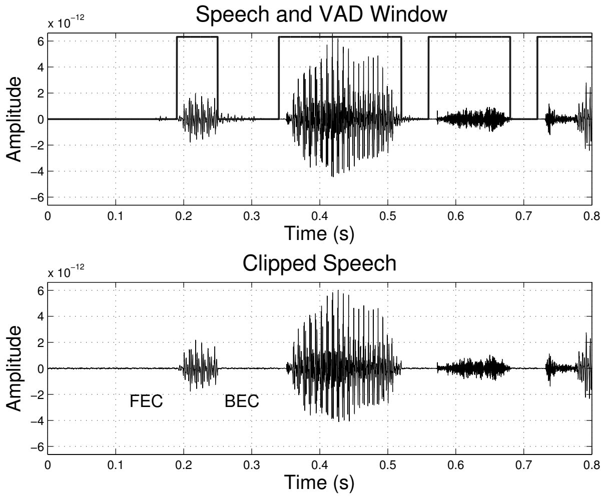

Figure 1: Example of original signal with VAD window with front end clipping (FEC) and back end clipping (BEC) illustrated on clipped signal below.

2. Background

Voice activity detection and line echo cancellation are standard modules employed in VoIP implementations. They provide a twofold benefit by improving the bandwidth utilisation by removing unnecessary noise transmission, and by removing unwanted echo and background noise from conversations. While line echo cancellation performs a different task to VAD, the failure mode is common between the two modules, with active speech frames being incorrectly marked as inactive and clipped. An example of this is shown in Fig. 1. The performance of VAD systems can vary significantly depending on the SNR of speech, type of environmental noise, and on the implementation [12, 13]. The location of the clipping within a talkspurt can vary and is usually broken down into three categories: front end clipping (FEC), middle speech clipping (MSC) and back end clipping (BEC) depending on the part of speech clipped. Front and back end clipping are illustrated in Fig. 1. FEC is the most noticeable from a speech quality perspective [11, 14]. An argument in provided by Gruber et al. [15] that mid-speech burst clipping can impact perceived quality more in longer passages of speech. This would be relevant in VoIP conversations but was not tested in the current work.

Ding et al. [11] proposed a non-intrusive metric to predict temporal clipping. The algorithm does not require access to the reference signal but it does assume prior knowledge of the clipping statistics for the degraded speech and proposes the used of method described in [16] to obtain a VAD estimate. This may not be a practical approach in a non-hetrogeneous VoIP scenario with a range of modules that can potentially introduce varying amounts of temporal clipping.

Detecting temporal clipping using a full reference speech quality metric is easily achieved as the unclipped reference is available. Metrics such as POLQA also predict the impact that clipping will have on the perceived speech quality. In a non-intrusive, no-reference environment, prediction of the impact of temporal clipping on speech quality is more challenging. The model must distinguish between natural silences or pauses and clipped frames. It needs to take into account that suppression of long segments of inactive speech will not impact the perceived quality, but a small amount of clipping at the front end of a talkspurt can noticeably mask speech and impact the perceived quality for the listener. There is not necessarily a direct correlation between the efficiency of a VAD algorithm and the effect on speech quality – classifying a silence frame as active would reduce the VAD performance in terms of bandwidth efficiency but would not impact speech quality while the reverse is not true.

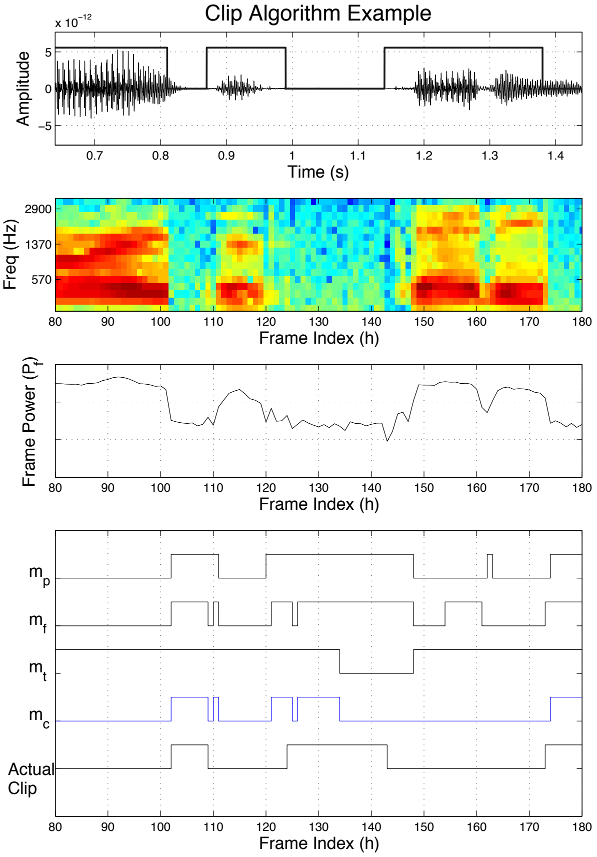

Figure 2: Temporal clipping detector algorithm example. The top pane shows a segment of speech with the VAD window marked. The next pane shows the STFT spectrogram with 15 critical band frequency bands and 8 ms frames. The third pane shows the a plot of the mean power per frame P_f. The bottom pane shows the three binary masks used to produce the clip mask, m_c, which is shown in blue. The actual clip mask at the bottom, which is the inverse of the VAD in the top pane. It can be seen that m_c does not follow the actual VAD mask directly but is likely to oscillate when clipping occurs.

3. Temporal Clipping Detection

The algorithm uses a short-term Fourier Transform (STFT) spectrogram of the test signal to measure changes in the mean frame power. An example is shown in Fig. 2. A STFT is created using critical bands between 150 and 3,400 Hz for narrowband speech, yielding 16 frequency bands. A 128 sample, 50% overlap Hanning window is used for signals with an 8 kHz sampling rate, giving a frame period $T_f = 8$ ms. For a signal N frames in length, the mean power $P_f[h]$ for frame, h, is calculated across the 16 frequency bands as:

$$P_f[h] = \frac{1}{16}\sum_{k=1}^{16} P[h,k] \tag{1}$$

A power mask is calculated as

$$m_p[h] = \begin{cases} 0 & \text{if } P_f[h] > \bar{P} \\ 1 & \text{if } P_f[h] \leq \bar{P} \end{cases} \tag{2}$$

where $\bar{P}$ is the average power over N frames. A frame gradient mask, $m_f[h]$, between the mean power of the high and low frequency bands is also computed using

$$y[h] = \frac{\sum_{k=1}^{3} P[h,k]}{\sum_{k=13}^{15} P[h,k]} \tag{3}$$

$$m_f[h] = \begin{cases} 1 & \text{if } y[h] > \bar{y} \\ 0 & \text{if } y[h] \leq \bar{y} \end{cases} \tag{4}$$

where $\bar{y}$ is the mean value of y over N frames. The power mask, m_p, is then used to calculate a "rough" voice activity activity mask, i.e one that detects talkspurts and masks larger silences but ignores short pauses within speech. This is needed to remove longer silence sections from the overall sample as the model calculates clipping as a rate with respect to the active speech length. This talkspurt mask, m_t is computed using a moving average filter to ensure short inactive sections frames at the onset of VAD are included. The 120 ms moving average is computed over J frames ($J = 15$ for 8 ms frames) as

$$q[h] = \frac{1}{J}\sum_{i=0}^{J} m_p[h-i] \tag{5}$$

$$m_t[h] = \begin{cases} 0 & \text{if } q[h] > \bar{P} \\ 1 & \text{if } q[h] \leq \bar{P} \end{cases} \tag{6}$$

To calculate the temporal clipping mask, m_c, a logical AND of m_f, m_p and m_t is computed

$$m_c = m_f \otimes m_p \otimes m_t \tag{7}$$

Finally, the estimate of temporal clipping , T_c, is computed as the number of zero crossings per second. This is computed from m_c when offset by $c_o = -0.5$, divided by the number of active samples in m_t times the frame period T_f as

$$z[h] = \begin{cases} 1 & \text{if } (m_c[h]-c_o)(m_c[h-1]-c_o) \leq 0 \\ 0 & \text{if } (m_c[h]-c_o)(m_c[h-1]-c_o) > 0 \end{cases} \tag{8}$$

$$T_c = \frac{\sum_{h=1}^{N} z[h]}{T_f \sum_{h=1}^{N} m_t[h]}. \tag{9}$$

4. Simulation and Model Evaluation

Two sets of speech samples were used to test the algorithm. The first used speech samples from the TIMIT database [?] and the second the IEEE database [18]. In both cases each speech sample contained a single talkspurt of 3–3.6 s in length and greater than 90% active speech. The TIMIT samples were chosen to match the variety of samples used by Ding et al. [11]. using samples from all of the eight dialect regions. Their preprocessing methodology was followed and the speech samples were downsampled to 8000 Hz, and level adjusted to -26 dBov using ITU-T Rec. P.56 level adjustment. The second test used 30 samples from the IEEE speech corpus [18]. Ten sentences from three speakers were used, each of approximately 3 seconds in duration.

As per Ding et al., an energy-based VAD algorithm was used to create temporally clipped speech. The algorithm segments the speech into frames and uses a threshold energy to determine if the frame is speech active or inactive. Four frame sizes were tested, 5, 10, 20, and 30 ms. For each frame size 9 thresholds were tested incrementing in 3 dB steps from 6 to 30 dB. The VAD algorithm replaced inactive frames with narrowband 30 dB pink noise.

This yielded 30 sentences with 36 temporally clipped samples per sentence. In addition, the reference sentences without any clipping were tested. As a baseline, the detector was also tested with choppy speech to validate that it was not susceptible to other type of similar VoIP degradations. Four frame sizes were tested, 5, 10, 20, and 30 ms. For each frame size 9 chop rates were tested incrementing in 1 Hz steps from 1 to 9 frames lost per second. The chopped frames were periodically spaced and replaced with zeros, i.e. without packet loss concealment.

The temporal clipping detector was used to estimate the temporal clipping in a no-reference manner. The clipped speech samples were also compared to the reference speech for each degraded sample using POLQA. The objective speech quality results were compared to the clipping detection algorithm results to assess its performance.

Ding et al. used PESQ as their benchmark quality metric. In this work, results were compared against POLQA as it has superseded PESQ as the ITU-T recommended speech quality metric.

5. Results and Discussion

The results are presented in Fig. 3 for the TIMIT and Fig. 4 IEEE tests. Examining the clip detection results for TIMIT first, the left pane of Fig. 3 plots the clip detected on the y-axis against the VAD threshold in dB on the x-axis. The results in red show the four frame sizes with 95% confidence intervals. It can be seen that the VAD energy threshold was the dominant factor and the size of the VAD frame was not a significant factor in the amount of clipping detected. Looking at the results from POLQA, it can be seen that the frame size did not significantly alter the MOS-LQO (Listener Quality Objective) quality predictions. The results for clip being detected fall to a level indistinguishable from the clean reference for the higher VAD thresholds but the POLQA results confirm that at such thresholds the clipping has a very small impact on perceived quality. In the right hand pane of Fig. 3, the results from the proposed model are compared to those from POLQA. A strong correlation is evident along the diagonal of the scatter, however there is a spread between the results for different frame lengths. This result was interesting as it was not evident in the IEEE results.

The tests with the IEEE database yielded similar results to those obtained with TIMIT. Although the absolute results differed (e.g. the reference scoring a clip level of 27 versus 25 for the TIMIT test), the results obtained from POLQA also differed, pointed towards a small difference in the impact of the VAD clipping on quality between databases. Overall, the scatter plots show a promising mapping between the clip scores and the full reference quality predictions from POLQA.

In both cases, the test with chop data was predicted to have a clip score within the same bounds as the reference samples, showing that the clip detector is not susceptible to other tran-

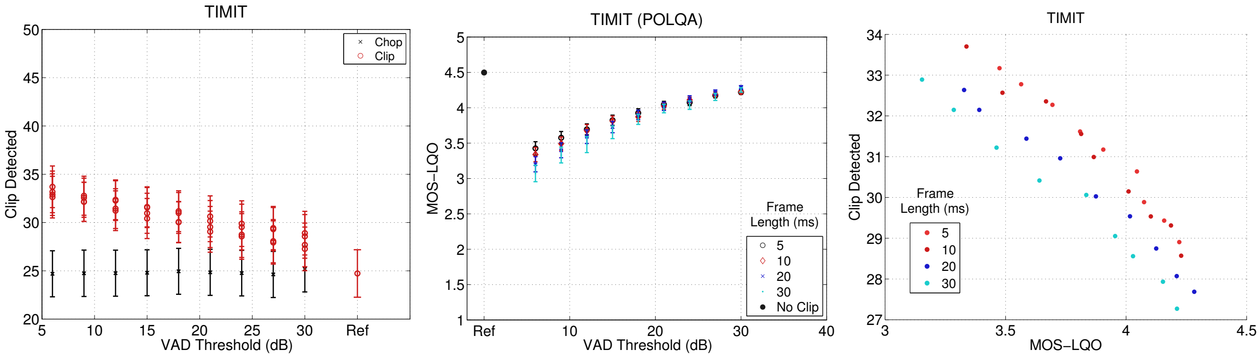

Figure 3: Results. Left: Results for TIMIT dataset. Middle: POLQA. Right: Comparison of results with POLQA.

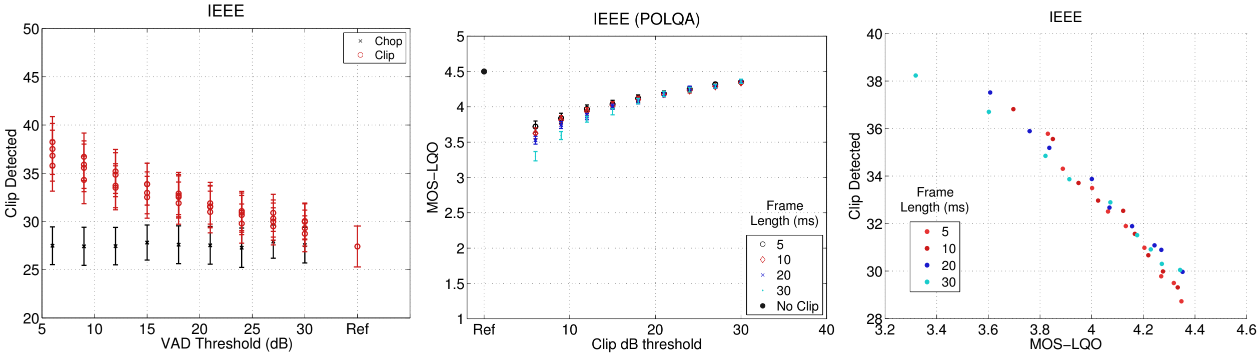

Figure 4: Results. Left: Results for IEEE dataset. Middle: POLQA. Right: Comparison of results with POLQA.

sient degradations impacting the signal across frequency bands.

The algorithm measures the amount of clipping activity rather than a ratio of the duration of speech clipped. A long clip of silence is not necessarily a bad thing from a quality perspective while clipped frames adjacent to active frames have an impact on quality. The tests were carried out using speech samples where the amount of active speech in the speech samples was kept to a similar ratio although the length of the samples differ slightly. Other than the lengths the main difference was the variety of speakers with the IEEE test containing 3 different speakers and the TIMIT test containing 14 (7 male and 7 female). Two other tests, the results presented in Fig. 5, were carried out using data from the ITU P.Sup 23 dataset [19]. Using 30 single talkspurt samples with one speaker per test, similar results were obtained. This confirmed that the results are not speaker, or utterance dependant.

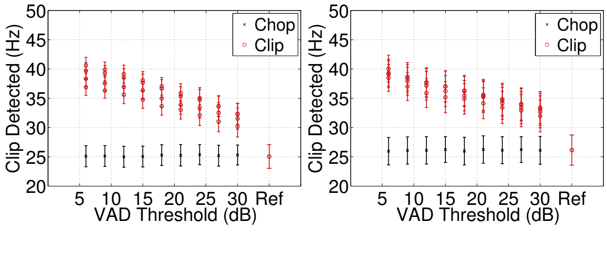

Figure 5: Results for ITU-T P.Sup 23 Database “O”. 30 single talkspurts from English male speakers 1 and 2.

6. Conclusions

This paper has presented a new algorithm for temporal clipping detection. The algorithm will form part of a modular speech quality estimation system. The detected clipping results demonstrate that the algorithm can efficiently predict temporal clipping in speech and correlates well with the full reference quality predictions from POLQA. The algorithm is no-reference and required no knowledge of clipping statistics. The tests here are performed on short (ca. 3 seconds) sentences. Further testing will be required to demonstrate whether the quality measure will transfer well to conversational situations. As observed in work by Gruber [15], quality degradations caused by repeated mid-word clipping may have a greater cumulative quality impact that can be captured in tests on short sentences. As observed by Moller et al. [4] the lack of publicly available data to develop and test speech quality metrics makes progress slow. Temporal clipping is common in VoIP and this detector could be easily deployed as a monitoring tool in a realtime system, or combined with other components as part of a more general no-reference speech quality model. Work by the authors in ongoing to develop additional components.

7. References

[1] "Coding of speech at 8 kbit/s using conjugate structure algebraic-code-excited linear-prediction (CS-ACELP)," Int. Telecomm. Union, Geneva, CH, ITU-T Rec. G.729, 1996.

[2] Google, "WebRTC NetEQ Overview," http://www.webrtc.org/reference/architecture.

[3] A. Raake, *Speech Quality of VoIP – Assessment and Prediction*. Wiley, 2006.

[4] S. Moller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: Models and trends," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 18–28, 2011.

[5] ITU, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecomm. Union, Geneva, CH, ITU-T Rec. P.862, 2001.

[6] ——, "Perceptual objective listening quality assessment," Int. Telecomm. Union, Geneva, CH, ITU-T Rec. P.863, 2011.

[7] ——, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Int. Telecomm. Union, Geneva, CH, ITU-T Rec. P.563, 2004.

[8] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1948–1956, 2006.

[9] A. ATIS, "0100005-2006: auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality," *American National Standards Institute*, 2006.

[10] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1935–1947, 2006.

[11] L. Ding, A. Radwan, M. El-Hennawey, and R. Goubran, "Measurement of the effects of temporal clipping on speech quality," *Instrumentation and Measurement, IEEE Transactions on*, vol. 55, no. 4, pp. 1197–1203, 2006.

[12] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 478–482, 2000.

[13] V. R. Prasad, A. Sangwan, H. Jamadagni, M. Chiranth, and R. Sah, "Comparison of voice activity detection algorithms for voip," in *Seventh International Symposium on Computers and Communications*. ISCC 2002, IEEE, 2002.

[14] N. Jayant and S. Christensen, "Effects of packet losses in waveform coded speech and improvements due to an odd-even sample-interpolation procedure," *Communications, IEEE Transactions on*, vol. 29, no. 2, pp. 101–109, 1981.

[15] J. Gruber and L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems," *Communications, IEEE Transactions on*, vol. 33, no. 8, pp. 801–808, 1985.

[16] A. Radwan, "Non-intrusive speech quality assessment in VoIP," *M.S. Thesis, Carleton University, Ottawa, ON, Canada*, 2003.

[17] U. D. C. DARPA, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," *NIST Speech Disc 1-1.1*, 1990.

[18] IEEE, "IEEE recommended practice for speech quality measurements," *Audio and Electroacoustics, IEEE Transactions on*, vol. 17, no. 3, pp. 225–246, Sep 1969.

[19] ITU, "ITU-T coded-speech database," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.Sup23, 1998.