A PITCH SYNCHRONOUS AUDITORY DOMAIN ANALYSIS TECHNIQUE FOR NON-INTRUSIVE AUTOMATED SPEECH DISTORTION IDENTIFCATION

Wei Wei and Liam Kilmartin

Communication and Signal Processing Research Unit
Department of Electrical Engineering
National University of Ireland, Galway,
Ireland
Email: wei.wei@nuigalway.ie liam.kilmartin@nuigalway.ie

ABSTRACT

This paper outlines research focused on the development of a non-intrusive methodology for the identification of speech distortions in packet-based transmission systems, e.g., voice-over-IP (VoIP) networks. In this field, speech quality degradation due to packet loss is of most interest. Methods presented in the paper focus on the localisation of abnormalities of changes in a sensation surface based on pitch perception techniques. Labeled pitch periods are initially aligned in length using a sample interpolation and power adjustment techniques. Subsequently, they are mapped into an auditory space similar to that used in PSQM/PESQ algorithms. Normal pitch period chains exhibit significantly different behaviour in the transformed perception space when compared with pitch chains containing a pitch distortion effect. Results are presented for initial tests which have been carried out to evaluate the performance of this algorithm in identifying distortions introduced by packet loss replacement algorithms.

1. INTRODUCTION

In packet-based speech transmission systems, such as VoIP, VoATM (voice-over-ATM), the sources of speech quality degradations differ from those typically found in traditional circuit-switched networks. The classical factors, e.g., noise in channel, filtering distortions, do not play a dominant role in speech quality degradation for such voice-over-packet networks. In contrast, packet loss due to traffic congestion in the case of a heavy network load has become the most dominant factor among channel related distortions. Since these lost packets cannot be restored perfectly in the receiver, many algorithms have been proposed to conceal the perceptual effect due to the packet loss replacement algorithm. Among these algorithms, replacing lost packets with a filtered version of the "last" correctly received packets is one of the most popular and simplest methods. Such an approach will typically create an perceivable distortion in the resultant speech signal.

The aim of non-intrusive speech quality evaluation techniques is to estimate the effects of distortions without the need for a reference (or distortion-free) speech signal. Techniques used in non-intrusive speech quality assessment can be classified roughly into three classes:

- Generation-based model - Speech abnormalities are identified using a model based on the speech production system (human speech organs), such as vocal tract areas as in [3],
- Perception-based model - Speech abnormalities are identified by analysis of the speech waveform in an auditory domain. Many auditory models with neural network processing or Computational Auditory Scene Analysis (CASA) may be classified into this class [7],
- Generic parametric models - Speech waveforms are parameterised and analysed using a variety of non-perceptual transformation techniques, e.g. HMM.

Research presented in [3] outlines a vocal tract model which uses a technique based on pitch period distortion identification as a proposed front end for an automated speech distortion identification system. The overall recognition capabilities of such a system is highly dependent upon the accuracy of the pitch labeling algorithm used prior to parameterisation of the voiced speech segments into vocal tract cross sectional area vectors. When the accuracy of the pitch-labeling algorithm begins to decrease (e.g. due to co-articulation, noise and/or distortion effects), the correct identification rate of such a system will naturally decreases. In [4], two output-based objective speech quality measures (OBM1 / OBM2, OBM—Output Based Measurement) based on

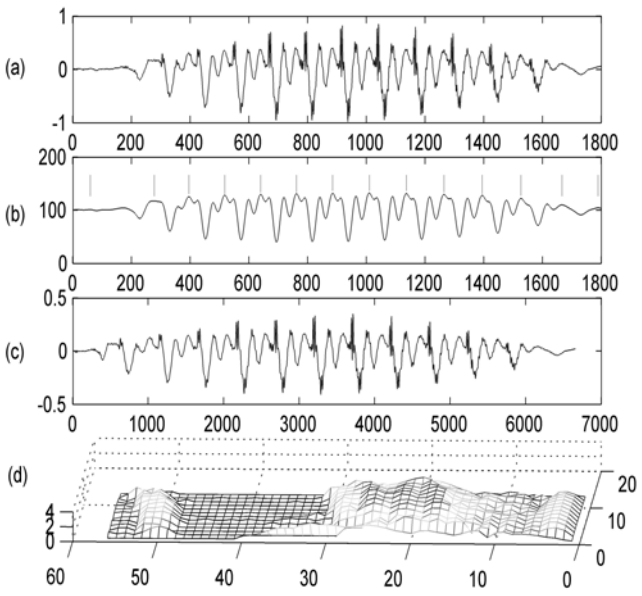

Figure 1 – Pitch synchronous auditory transformation
(a) Original utterance, "he" from a male speaker
(b) Pitch labelling for that utterance
(c) Normalisation of pitch periods
(d) Auditory transformation of normalized pitches

spectrogram comparisons between consecutive analytical frames are outlined. While this research outlined an approach to estimate the effect of a distortion on a speech signal, it does not provide a means of identifying the location of such distortion with any degree of accuracy (which is a desirable procedure in many applications).

In ITU-T recommendations P.861 [1], now superceded by recommendation P.862 [2], objective speech quality assessment methodologies PSQM (Perceptual Speech Quality Measurement) and PESQ (Perceptual Evaluation of Speech Quality), respectively, are outlined. While both techniques outline paradigms for "intrusive" speech quality estimation, the core auditory transformation used within both algorithm offers a flexible technique of observing speech distortions in an auditory domain. The research presented in this paper attempts to re-use this core auditory transformation even though it is applied to labeled pitch periods rather than for speech frames with a fixed-length.

The primary consideration of this study was to develop an algorithm which is capable of locating the presence of distortions due to typical VoIP packet loss replacement techniques. If sufficiently accurate and robust, this algorithm could then be used as the front-end processing stage for a non-intrusive speech quality evaluation system. In section 2, the auditory mapping for labeled pitch periods is introduced including a technique for pitch period normalisation in terms of both length and power. Section 3 discusses various observations that were made regarding the representation of distorted and distortion-free speech segments when transformed into an auditory perceptual space. Also outlined in this section are the resultant analysis procedure and rules for identifying the presence and location of speech distortions in the pitch synchronised speech. Section 4 presents the results obtained for initial testing of the overall proposed algorithm on selected speakers' voices from the ITU-T P.23 supplement database. Section 5 presents some conclusions drawn from these results along with suggestions for improvement of the method (which are currently under investigation).

2. PITCH SYNCHRONOUS AUDITORY MAPPING

Pitch synchronous parameterisation of speech enables analysis of speech at a pitch period level. The final accuracy of the pitch synchronisation procedure is dependent upon the pitch labeling methodology as well as the presence of any distortion artifacts in the speech segment under analysis.

2.1. Pitch Period Labeling

The general approach of the pitch period labeling method employed here was suggested in [5]. An original utterance and its labeled pitch periods can be seen in figures 1 (a) and 1 (b).

The features of the labeling procedure are:

- The labels of pitch periods are independent of speech degradation, except for a few severe distortion conditions, e.g., SNR<10dB, packet loss rate > 5%, etc. as in [5],
- Pitch period labels are selected from the peaks in the processed utterances with two consecutive labels denoting a pitch period. The waveforms of "clean" pitch periods within an utterance are similar in some parameters, such as slowly changing number of samples (jitter of f_0, fundamental frequency), slowly evolving power (shimmer), etc., while waveforms of "abnormal" pitch periods may vary in any way,
- Pitch period length is limited to the range of 3 ms to 16 ms.

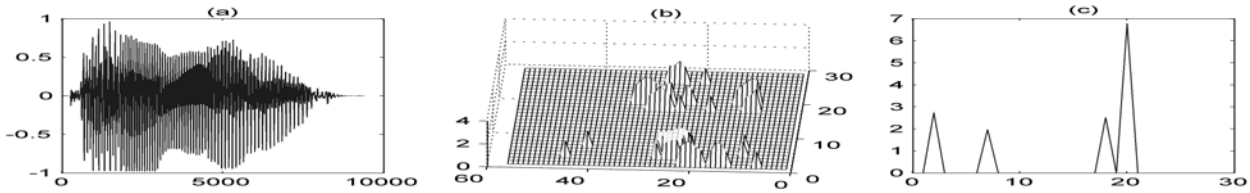

Figure 2 - Auditory domain smoothness assessment and simplification in distortion free speech
(a) Original speech (male speaker voice "Galway"),
(b) Simplified smoothness metric across critical band *k* and block number *n* ,
(c) Simplified smoothness in block number *n*, $d_1(n)$.

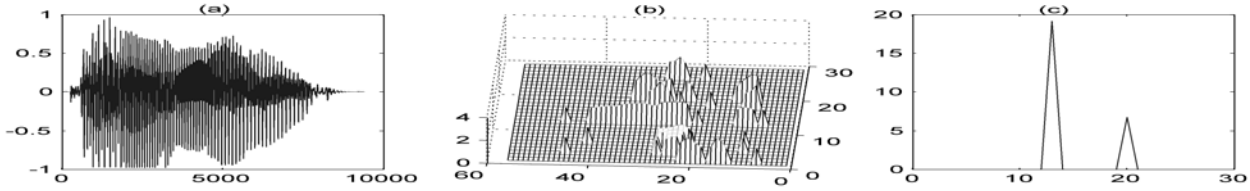

Figure 3 - Auditory domain smoothness assessment and simplification in speech containing distortion
(a) Distorted speech (simulated packet loss replacement occurs between samples 3201 and 3520),
(b) Simplified smoothness across critical band *k* and block number *n*, $d_0(k,n)$,
(c) Simplified smoothness in block number *n*, $d_1(n)$.

2.2. Normalization of Pitch Periods

Labelled pitch periods differ from each other in many ways, such as length, waveform, and mean power. These differences potentially prevent them from being mapped into, and subsequently accurately compared, in a perceptual auditory space. Thus normalization for individual pitch period is necessary. In this research, the normalisation procedure includes:

- Pitch period length normalisation. All the pitch periods were interpolated to 512 samples [frame size used by PSQM analysis]
- Band pass filtering. A frequency domain band-pass filtering is carried out on the waveform with the mean power of the original waveform being conserved.
- The amplitude of the complete sentence under analysis is normalized.

Figure 1 (c) is the normalized output of the pitch periods in figure 1 (d). While it may seem that this form of normalization is somewhat arbitrary, it does result in an overall algorithm with a better performance and offering superior robustness.

2.3. Transformation of Pitch synchronous Speech into Auditory Domain

The normalized pitch synchronised speech segments is then transformed into a perceptual auditory space (namely the Bark-based space outlined in [1]). Normalized pitch waveforms are transformed into this sensation space in tandem. The dimension of the auditory space is 56 critical bands and the subjective perception of the signal is reflected in the loudness distribution along these bands. An illustration of the transform output is shown in figure 1 (d).

3. SMOOOTHNESS DETECTION WITHIN PITCH PERIOD GROUP

A substantial amount of time was spent analysing various perceptual transformations of pitch synchronised speech segments and copies of these utterances when a simulated distortion (i.e. packet loss replacement artifact) was introduced. Inspired by the assessment in [6] of the subjectivity of perceptually weighted audible errors, a visual auditory domain "smoothness" is suggested here for the detection of speech "abnormalities". Figure 2 and Figure 3 illustrate the proposed procedure, where auditory domain smoothness of both distortion free and distorted utterances were observed and simplified by means of pitch period grouping.

3.1 Pitch Period Grouping

The size of pitch block here refers to the number of consecutive individual pitch periods that are grouped together for the purpose of subsequent analysis. The procedure of combining together a group of individual pitch periods and analyzing these as a single block (rather than analyzing individual pitch periods) takes into account the following considerations:

- The representations of individual distortion-free pitch periods in this auditory space can be quite diverse. The boundary is not only due to the 56 critical band expressions as in PSQM [1], but also due to the dynamic range within each band. The properties of a typical "normal" individual pitch period are difficult to identify in this auditory space.
- Distortions due to lost packet replacement techniques generate "abnormal" pitch periods which can be perceived by a listener but which are difficult to identify when analyzing individual pitch cycles,
- Comparisons between consecutive pitch periods can show quite similar or quite diverse properties (for both distortion-free and distorted speech utterances).

In this study, the pitch period block size was set to 3 (with no overlapping of consecutively analysed pitch period blocks). For female speakers, this typically corresponds to a time duration of about 12 ms to 18 ms, while for male speakers, it is the equivalent of between 15 ms and 30 ms. Experiments supported this choice of block size as being relatively optimal and provided a good balance between a longer block size (which would result in poorer resolution in identifying the location of distortion in a speech utterance) and shorter block sizes (with the associated problems identified previously).

3.2 Smoothness Analysis within Pitch Blocks

Inspired by the definition of OBM definition in [3], the smoothness within a certain pitch block n is defined according to following formulation:

$$d_0(k,n) = \max(CB(k)) - \min(CB(k))$$

where

k: critical band number in PSQM,

n: pitch block number, dependent on the length of analysed utterance.

$CB(k)$: k^{th} component of the (56 Bark components) critical band transformation of the speech signal.

The selection of the range of k values to be used (namely k_1 and k_2) was determined by trying to optimise the difference between distortion-free pitch period blocks and pitch period blocks containing a distortion artifact. In particular, observation noted during the analysis of the simulated "distorted" speech database, suggested that there are not significant differences between $CB(k)$ components at lower frequencies. For this reason, it decided that these components would not be of significant use in the process of distortion identification and hence it was decided to use the range of k_1=7 to k_2=56 in the calculation of the surface $d_0(k,n)$.

3.3 Simplification of "Smoothness" Metric within Pitch Period Blocks

While the surface defined by $d_0(k,n)$ reflected some desired features in the observation space, initial tests showed it to be somewhat lacking as an adequate a parameterisation technique for identifying the presence of distortions in a speech waveform. Thus, in order to enhance its reliability in abnormality detection, all $d_0(k,n)$ values in the same utterance were adjusted as follows:

- Initialise threshold value THR
$$THR = \max(d_0(k,n))/2$$
- Re-set the first pitch period block value to 0
$$d_0(k,1) = 0$$
This is due to the fact that auditory changes in first pitch block often is bigger and may mask the appearance of a distortion artifact.
- Maintain $d_0(k,n)$ if it is bigger than THR, otherwise set to 0
$$d_0(k,n) = \begin{cases} d_0(k,n) & \text{if } d_0(k,n) > THR \\ 0 & \text{if } d_0(k,n) \le THR \end{cases}$$
- Maintain $d_0(k,n)$ if pitch block n has sufficient non-zero components (>10% of critical bands), otherwise set it to 0.

The graphs in figure 2(b) and figure 3(b) show this simplified comparison technique of pitch period blocks for both distortion free speech and distorted speech. A packet replacement distortion was simulated in the sample range 3201-3520. This distortion region corresponds to pitch period blocks 13 and14, which can be distinctly identified as a ridge in the auditory space mapping.

3.4 Simplification of Smoothness Metric across Pitch Period Blocks

While the steps outlined thus far seemed to provide a representation of speech suitable for distortion identification, further experiments with additional

speakers and speaking conditions highlighted a number of additional issues. These problems included:

- Distortion-free utterances and distorted utterances both have non-zero $d_0(k,n)$ data sets. It would be more desirable that an "all-zero" $d_0(k,n)$ is indicative of a distortion-free utterance.
- Some distorted utterances have spurious non-zero $d_0(k,n)$ vectors. It would be desirable to decrease the number of such spurious indications as much as possible.
- The use of the "PSQM-like" 56 critical band representation for speech may be unduly complex.

To deal with these problems, distorted pitch periods were modelled as a summation of clean pitch periods and a related white noise signal in auditory space. The noise is assumed to have same temporal length and a similar spectrum to the distortion free pitch periods, while its power is dependent upon the severity of the distortion artifact which is present. The following features were observed when comparing the distortion-free pitch period blocks with their noise "shadow" blocks (which contained the same mean power):

- *CB(k)_clean* changes quickly with *k*, *CB(k)_noise* is almost constant with *k*,
- *CB(k)_clean*>*CB(k)_noise* in lower bands(k<7), *CB(k)_clean*<*CB(k)_noise* in higher bands (k>40),
- Non-zero $d_0(k,n)\_clean$ components are randomly distributed in pitch period blocks and tends to have smaller values while non-zero components in $d_0(k,n)\_noise$ tends to be concentrated in the higher frequency bands and tends to have larger values.

While further features were noted from these comparisons, it is not easy to combine all of these observations effectively. The final representation of the "smoothness" of this auditory domain representation of a pitch period block is a weighted scalar defined by:

$$w(n) = \frac{1}{(k_2 - k_1)} \sum_k \partial_k$$

$$\text{where } \partial_k = \begin{cases} 1 \text{ if } d_0(k,n) \neq 0 \\ 0 \text{ if } d_0(k,n) = 0 \end{cases}$$

$$d_1(n) = w(n) \sum_k d_0(k,n)$$

All pitch periods within the same pitch block *n* are labelled with the same $d_1(n)$, which is an indication of the probability that the pitch period block contains a distortion artifact.

4. ALGORITHM EVALUATION

The ability of the proposed algorithm to detect the presence of distortions in voiced speech utterances was evaluated in initial tests using voice recordings of both male and female speaker selected from the ITU P.23 supplement database.

Since the proposed algorithm is based on the analysis of pitch period blocks, the following constraints were used during these tests:

- Utterance length > 150ms (including at least 9 pitches),
- Absolute threshold of $d_1(n)$ was set at a value of 4. For any $d_1(n)$<4, the pitch block *n* is regarded as normal. This threshold setting results in the filtering out of most false identifications in distortion-free utterances.,
- Removal of insignificant smoothness indications:

$$d_1(n) = \begin{cases} d_1(n) \text{ if } d_1(n) > \max(d_1(n))/2 \\ 0 \quad \text{if } d_1(n) \leq \max(d_1(n))/2 \end{cases}$$

- Distorted regions, if represented with non-zero $d_1(n)$ values, are interpreted as identifying the presence of a speech-distortion.

The performance of the algorithm is quantified using the two metrics *IDR* and *IDX*, which are defined as follows:

- *IDR(%)*: percentage of correctly identified pitch periods containing a distortion artifact,
- *IDX(%)*: percentage of falsely identified distorted pitch periods.

Table 1 shows a sample set of results (for two male and two female speakers) from the initial tests carried out. the parameters of voice recordings under test and the test result.

SPK	PNo	SNo	Rno	XNo	IDR	IDX
M01	1092	123	108	63	87.8	6.5
M02	1631	117	102	54	87.2	3.6
F01	1926	199	168	37	84.4	2.1
F02	1780	172	144	54	83.7	3.4
Mean	6429	611	522	208	85.4	3.6

Table 1 – Summary of Evaluation Test Results

SPK=speaker, M01=male speaker1, F01=female speaker1
Pno : Total number of pitch periods in test,
Sno : Number of pitch periods containing a distortion artifact,

Rno : Number of correctly classified "distorted" pitch cycles,
Xno : Number of falsely identified "distortion-free" pitch cycles,
IDR and IDX : Percentage (%) of Rno and Xno, respectively.

5. CONCLUSIONS AND DISCUSSION

The purpose of this research is to achieve a means of identifying the presence and location of abnormalities in speech (Specifically those introduced by VoIP packet loss replacement algorithms) in a non-intrusive manner. These distortions can be readily identified (and their impact on speech quality subsequently estimated) when a reference speech signal is available. This paper has outlined a technique suitable for non-intrusive speech distortion identification specifically applicable to distortion resulting from packet loss replacement algorithms typically used in VoIP networks.

While the recognition rates obtained in initial tests of the algorithm are by no means ideal, the mean recognition rates of IDR=85.4% and *IDX*=3.6% are reasonably promising. Some analysis has been carried out on the main reasons for the failure of the algorithm to operate correct. The main contributory factors are due to:

- Uneven speech characteristics in the utterance. This includes weaker transitional area in the middle and at the boundaries of utterances. Such natural changes in the speech waveform appear to be very similar to the characteristics of distortion artifacts in the auditory domain being used,
- The description of the auditory domain's "smoothness" provided by $d_0(k,n)$ does not generalise (to other speakers and in some cases other utterances) as well as expected.

Current research is focused on achieving better recognition rates for the algorithm. In particular, current work is focusing on considering the following issues:

- Impact of failures of the pitch labeling algorithm on the overall algorithm performance,
- Pitch period block size optimization.
- Improving the methodology for determining the auditory transforms $d_0(k,n)$ and $d_1(n)$. Even though the present method for calculating $d_0(k,n)$ and $d_1(n)$ are quite effective, they intuitively provide less connection with subjectivity test results,
- The current algorithm treats individual pitch period blocks as isolated entities. The issue of the relationship between pitch period blocks within individual utterances and adjacent utterances needs further investigation,
- A pitch period feature database built using an analysis of a sample set of a few speakers, is expected to be used as a generalised "speaker independent pitch feature database". Such a database will be helpful to achieve a standardised pitch period analysis procedure which is applicable to *any* speaker.

REFERENCES

[1] ITU-T Rec. P.861, "Objective quality measurement of telephone-band (300-3400 Hz) speech codecs", ITU-T, February 1998

[2] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs", ITU-T, February 2001

[3] P. Gray, M.P. Hollier and R.E. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," IEE Proceedings--Vision, Image Signal Processing , pp493--501 Vol. 147, No. 6. December 2000.

[4] O.C. Au, K.H. Lam, "A Novel output-based objective speech quality measure for wireless communication," Proceedings of ICSP '98, pp666-- 669,

[5] W. Wei and L. Kilmartin, "Segmental optimization of automated pitch labeling for distorted speech signals, " Proceedings of ISSC 2003, Limerick , Ireland, July 1-2 (in press)

[6] M.P. Hollier, M.O. Hawksford, D.R. Guard, "Algorithm for assessing the subjectivity of perceptually weighted audible errors ", Journal of Audio Engineering Society, Vol. 43, No.12, December 1995

[7] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'99), vol. 2, pp. 929-932, Phoenix, Arizona, March 15-19, 1999.