Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Tolerance and sensitivity of various parameters in the prediction of temporally localized distortions in degraded speech

Wenliang Lu (1) and D. Sen (1)

(1) School of Electrical Engineering and Telecommunication, University of New South Wales, Australia

PACS: 43.71.Gv

ABSTRACT

Objective speech quality measurements can be made more accurately and robustly by analyzing individual distortions of the afflicted speech signal. In previous papers [7, 8] "Salient Formant Points" (SFP) that are extracted from the output of a hydromechanical, physiologically motivated cochlear model have been used to predict the perceptibility of Temporally Localized Distortions (TLD). The feature represents areas of high energy vocal tract resonances resolved in the output of the cochlear model. TLDs include afflictions best described using words such as "Fluttery", "Babbly", "Harsh" and "Interrupted" and represent the highest variance amongst speech signals subjected to coding, network errors as well as environmental noise. In previous work [7, 9] we have reported the high correlation between the predicted and actual subjective Diagnostic Acceptability Measure (DAM) elementary perceptual quality (EPQ) scores that represent these distortions. In this paper we investigate the algorithm's tolerance to the alteration of various factors that affect the accuracy of the TLD prediction. The parameters investigated include misalignment between the original and degraded speech signals; inaccurate voiced/unvoiced decisions; inaccurate speech level normalization at the input to the cochlear model (which affects the output of the non-linear cochlear mode). Results are illustrated for each these factors.

INTRODUCTION

The accurate and objective measurement of speech quality has been the "Holy Grail" of speech processing. The current ITU standard for objective measurement of speech quality, P.862 (also known as the "Perceptual Evaluation of Speech Quality" (PESQ) algorithm [6]) - is often inappropriate, (as documented in the standard), for evaluating low bit-rate vocoders (below 4kbps) [6] as well as speech degraded by environmental conditions. These environmental degradations include babble/office noise and military vehicle noise which affect the signal before being subjected to coding. In realistic conditions clean (not affected by the environmental conditions) versions of those signals do not exist and cannot be provided for use as the reference signal to the PESQ algorithm. In addition, our own tests reveal that PESQ fails to predict the quality of low pass filtered speech ($f_c = 2kHz$) as well as speech degraded by narrow band noise (from $400Hz$ to $800Hz$). Even so, the PESQ algorithm betters earlier attempts at predicting MOS [1] - mainly attributed to a highly evolved Psychoacoustic Masking Model (PMM). The PMM is an attempt at modelling the linear component of what is a highly non-linear hydromechanics of the human cochlea.

A fundamental issue that affects the accuracy of speech quality evaluation is that both subjective Absolute Category Rating (ACR) testing and PESQ tries to evaluate the quality in one dimension. This is counter to evidence [4, 12, 14], that the perception of speech quality is multi-dimensional. This is addressed in the proprietary subjective testing method called Diagnostic Acceptability Measure (DAM) [2]. In DAM, subjects are asked to detect individual distortions such as "Babbly", "Interrupted", "High-frequency distortion", etc, before being amalgamated into a single score (CAE) that correlates extremely well with ACR scores. In previous research, we have employed the same "Divide and Conquer" strategy [12] to the objective evaluation of speech quality to measure Frequency Localized Distortions (FLD) [11] and Temporally Localized Distortions (TLD) [7, 8]. These have proven to be extremely effective and also been used to predict ACR scores more robustly than PESQ [13].

During the development of TLD measurement using the SFPs, some interesting characteristics of the SFPs have been revealed. These include a highly desirable "auto-alignment" property of the SFPs. Most intrusive objective speech quality measures spend considerable amount of computation attempting to align the reference signal with the distorted signal. This paper continues to investigate the auto-alignment property of the SFPs along with some other factors that affect the performance of TLD prediction using the SFPs.

This paper is organized as follows. The following section will describe the overall TLD measurement algorithm. The next section will discuss three experiments that examine the tolerance of various parameters of the algorithm that impact on the performance of TLD prediction.

FEATURE EXTRACTION FOR TEMPORALLY LOCALIZED DISTORTIONS

Temporally localized distortions in DAM are best described using "Babbling", "Fluttering", "Interruption", as well as a faster variation broadly perceived as "Harsh". As the names suggest, these distortions are localised in time - some appearing at a faster rate than others. The algorithm developed to predict their detectability is based on a few fundamental hypothesis.

The fist hypothesis is that a hydro-mechanical cochlear model (CM) which attempts to convert the speech signal into a domain that is closer to the perceptual domain - will do so more accurately than other existing methods (such as a linear Psychoacoustic Masking Model, or even short-term-Fourier-transforms). The second hypothesis is that humans evaluate speech quality mostly during voiced sections - which typically have longer durations and higher energy than other sections.

Perceptually salient features are extracted from the CM response of voiced sections to predict TLD distortions. The 2D Cochlear Model response across time $CM_p(t)$, at a single discrete place p (of arbitrary units), is a quasi-periodic waveform, with primary period T_c, dictated by the characteristic frequency $f_c = 1/T_c$, at place p. For voiced speech, a second mode of periodicity T_p can also be observed on the smooth low-passed envelope of the signal $e_p(t) = E\{CM_p(t)\}$. This periodicity is due to the pitch of the speaker and is independent of place p except for a slow evolution across space. These are shown for a typical voiced section in Fig. 1.

Due to causality, at place $p+1$, the envelope of the Cochlear Model response $e_{p+1}(t)$ will have evolved albeit slowly for voiced sections. The rate of evolution is a function of the amount of voicing, such that for highly voiced sections, this evolution is slow, whereas the rate is fast for unvoiced sections. The exact same argument can be made in the alternate dimension of looking at the Cochlear response as a function of place at discrete time t_0 and its evolution at $t_0 + 1$. It is necessary to track this evolution in both space and time dimensions since the envelope is evolving in both dimensions. Fig. 2 illustrates this evolution for a voiced section of speech by a 2D peak tracking algorithm.

We have adopted these peak tracks of the CM response as essential features that represent the rate of evolution of the response. It can be observed that the peak tracks are almost parallel when the rate of evolution is slow as is the case for voiced speech. This parallel structure is lost for unvoiced sections of speech and is shown in Fig. 3.

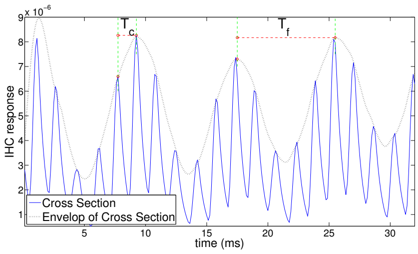

Figure 1: *Cross section through the CM response for voiced speech. Two types of periodicity, T_c and T_p, can be observed. T_c is given by the characteristic frequency of the place where the cross section is taken, while T_p is determined by fundamental frequency of this speech segment.*

The output of the cochlear model is two dimensional data across time and space. The sampling rate at the output is identical with the input speech signal while the spatial sampling is $0.0684 mm/sample$ such that there are 512 discrete points across the approximate $3.5cm$ length of the human BM. It is possible to convert between place and frequency using Greenwood's map [3] (at threshold levels).

The steps below describes an algorithm to track the two dimensional evolution of the cochlear response $CM_p(t)$ on a closed spatial region $p = [p_l, p_h]$ along the BM.

1. We start at the lowest boundary place p_l, which corre-

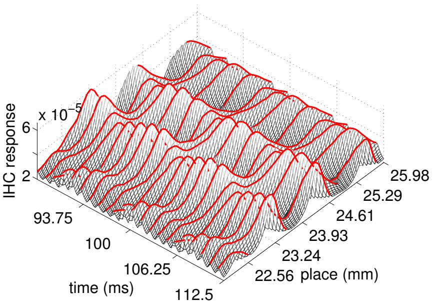

Figure 2: *Cochlear response as a function of time and place, with peak tracks for an voiced segment of speech (/o/). Dark lines indicate the peaks or crests of the response, and exhibit a regular, quasi-periodic structure which is also evidenced in Fig. 1.*

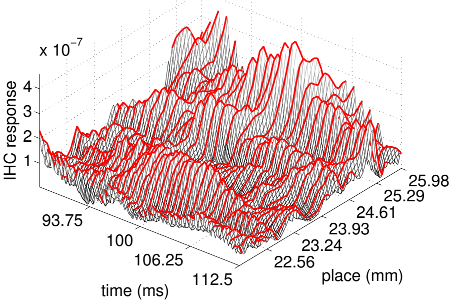

Figure 3: *Peak tracks from the cochlear response for an unvoiced segment of speech (/s/). The quasi periodic structure that appears in Fig. 2 is not present. Note, that the actual CM response is not plotted for reasons of clarity.*

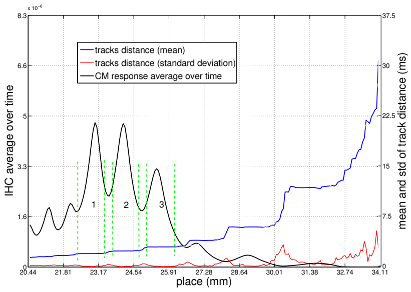

Figure 4: *Formant Places Determination. Track distance (in blue) levels out at some region, where its standard deviation is correspondingly lower than elsewhere. Only three regions with higher energy has been marked as Formant Places, as shown by 1, 2 and 3.*

sponds to the highest frequency in the region $[p_l, p_h]$. Find all local maxima along the time axis $CM_{p=p_l}(t)$, such that there are M_{p_l} peaks at time $t_k, k = 1, 2, \ldots, M_{p_l}$. The peaks are chosen such that at time t_k,the cochlear response $CM_{p_l}(t_k)$ satisfies the criteria that it is larger than the N neighbouring time samples, on either side of it, as follows: $CM_{p_l}(t_k) > CM_{p_l}(t_k - 1) > CM_{p_l}(t_k - 2) \cdots > CM_{p_l}(t_k - N$, and $CM_{p_l}(t_k) > CM_{p_l}(t_k + 1) > CM_{p_l}(t_k + 2) \cdots > CM_{p_l}(t_k + N)$. The value of N is a function of the temporal sampling rate and is empirically calculated to ensure the capture of salient features.

2. The process in Step 1 is repeated for each spatial point in the range $(p_l, p_h]$. The position of the peaks are stored in a matrix PT, such that $PT(p_c, k) = t_k, k = 1, 2, \cdots, M_{p_c}$. The size of the matrix is given by the maximum number of peaks at any place (i.e., $max(M_p)$).
3. The next step is to associate each peak with a track across time and place. To do this we look in a distinct neighborhood (i.e., $[t_{k,p-1} - t_{backword}, t_{k,p-1} + t_{forward}]$) of each peak position from the previous place, $p - 1$. If a peak is found within the above range, then it is considered to be part of the same track as the one at $t_{k,p-1}$. If more than one peak is found within that range, then the one closest to $t_{k,p-1}$ is chosen. If no peaks are found within that range, then it track is terminated at place $p - 1$ and no further search along this track is performed in the future. Due to causality, the peak tracks always move towards increasing time and place. For this reason, $t_{backward}$ can be small. It is important to account for any new tracks that originate at a higher place (i.e., was not at place $p - 1$) by ensuring that new peaks not associated with the previous place are not discarded but are stored for future tracking until they terminate.
4. Further post-processing involves checking to ensure that the track lengths are longer than a certain threshold. If not, these short tracks are discarded.
5. The final tracks are stored in a matrix $T(m, n)$ where each column describes a single track.

Example of the above steps is illustrated in Fig. 2 and 3. The continuous lines capture information on the evolution of the spectrum over time and space. During voiced speech, this evolution is slow and is characterised by peak tracks which do not change drastically over time and therefore take-on an almost parallel looking tracks across time and space.

Formant frequencies or vocal tract resonances are easily distinguishable in the 2D CM response. During voiced speech, they show up as distinct "peaks" or high energy regions in the CM response, as can be observed in Fig. 2. In the figure, the three formant frequencies can clearly be tracked over time and place. They appear at approximately $23.11mm$, $24.20mm$ and $25.57mm$ from the base of the BM, their positions changing slightly with time. These places correspond to approximately $4461Hz$, $3707Hz$ and $2911Hz$. Instead of referring to Formant frequencies, it is more appropriate to refer to these as Formant Places (FP), reflecting the association between each place along the length of the cochlea with a characteristic frequency.

The peak tracking algorithm described in the previous section tracks the FPs extremely accurately over time and place. This is one of the main reasons that the use of CM response is far superior than spectrogram, as the CM response reflects only the information that remains after non-linear cochlear processing. What is actually being tracked is the effect of the formants in the cochlea rather than the actual formants.

One of the important features of the Formants is their stationary nature over time and place. This can be observed on the CM response by the fact that the number of peaks remain unchanged for the duration of the voiced speech, as well as the fact that the peak-tracks are approximately parallel to each other (in the 2D projection across time and place) - especially in the regions of the Formant Places. This is demonstrated in Fig. 1.

The next step in our feature extraction is to focus on just the "Formant Places". This is facilitated by the observation that the average time difference between the peak tracks $\overline{\Delta_{t_p}} = \frac{1}{K-1}\sum_{k=2}^{K}(t_{p,k+1} - t_{p,k})$ (over the duration of the voiced section) is almost constant across the region of each Formant Place. This is shown in Fig. 4 which shows that in each of the three Formant Places, 1, 2 and 3, the $\overline{\Delta_{t_p}}$, shown by the blue line, is almost constant along the width of the each of three formant places. The standard deviation of the time difference, shown in red, is also shown to be low. Further, there is a conspicuous increase in the average time difference with increasing distance - such that the $\overline{\Delta_{t_p}}$ for region 1 is lower than the $\overline{\Delta_{t_p}}$ for region 2. This is a direct consequence of the fact that the number of peaks at any one places are lower with higher distance, reflecting the fact that the characteristic frequencies $1/T_c$ decreases with distance.

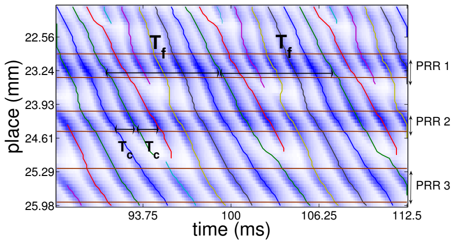

Figure 5: *Cochlear response with peak tracks for voiced speech /o/ on the time-place plane. The parallel structure between tracks can be observed at the FPs (between yellow straight lines). Also, the T_c and T_p in Fig. 1 is indicated here.*

By using a two pronged strategy of imposing an energy threshold such that only sections of the CM response above the threshold will be kept as well as using the graded characteristic of $\overline{\Delta_{t_p}}$, it is possible to concentrate only on the Formant Places, essentially discarding the rest of the CM response and associated peak tracks. The regions that were approximately kept after this stage are shown in Fig. 5 as the areas between the straight lines.

A characteristic of the peak tracks at the FP region is the fact that they are quasi-parallel on the time-place plane (much more so than in other regions). Corresponding tracks across period T_p, are also more similar in intensity than say neighbouring tracks. In a further attempt at reducing dimensionality, while keeping the most salient component of these tracks, we reduce each set of tracks in a single period T_p to a single point given by the "centre of mass" of the tracks in one period. Fig. 6 indicates the final result of this process. Fig. 6.(A) shows the extracted Salient Formant Points (SFP) in 3D space of time, place and IHC response. Fig. 6.(B) is a plot of the points showing the respective time they were extracted. A most notable feature is that the points extracted in this manner, for the two different systems are automatically synchronized - without the explicit requirement of the signals to be synchronized accurately at the input. Fig. 6.(C) shows that the points are lightly dispersed over place due to the different coding systems - as should be expected. Finally, Fig. 6.(D) shows the IHC response at each of the extracted points.

SB and SF are defined [2, 10] as "Babble" and "Fluttering" distortions respectively. From observation of systems which

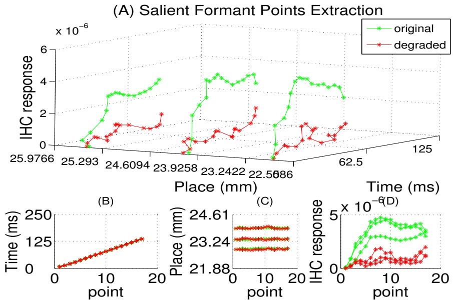

Figure 6: *Extracted Salient Formant Points. (A) illustrates both original (green) and distorted (red) SFPs as a function of time and place. (B), (C) and (D) shows the time, place and IHC response of the SFPs, respectively.*

have high SB and SF distortions, it can be deduced that they are highly influenced by temporally localised distortions. This is also reflected in the actual descriptions of these parameters ("interrupted" and "clipped" for example). This implies that the distortions are spread over frequency (or along the complete length of the BM/cochlea). However, this is further complicated by the fact that the CM response is not synchronized between the original and distorted signals - due to the fact that the CM is nonlinear and also a number of upsampling and downsampling steps that are carried out in the CM. This problem is alleviated by our use of the SFP feature which has the property of automatic synchronization, as shown in Fig 6.(B).

It is recognised that an actual sustained difference in the IHC response (between the original and coded speech) means little in terms of invoking a temporally localised distortion. Instead, a temporally localised distortion will introduce a highly fluctuating difference in the IHC responses. We hypothesize that this "jitter" or "trembling" is captured by the standard deviation of the difference in IHC responses, at the extracted SFPs, as shown in Equation 1. Also, as in our previous work [11], we only carry out this analysis in voiced areas of the speech signal with the hypothesis that speech quality is largely determined in voiced areas (whereas intelligibility is discriminated in unvoiced consonant areas) of the speech signal.

$$jitter = std(IHC(SFP_{ori}) - IHC(SFP_{dis}))|_{voiced} \qquad (1)$$

EXPERIMENTS

The above algorithm delivers excellent prediction of TLD perception. The purpose of this section is to analyse the performance of the algorithm when some basic assumptions are contradicted or errors are introduced. Specifically, we measure the performance when the original and reference signal are misaligned, the SFPs are analysed in unvoiced areas and when the speech signals are introduced into the cochlear model at inaccurate levels. The last issue is significant as the cochlear model is non-linear and we have assumed that an active speech level of $-26dBoV$ (as scaled by the ITU tool svdemo) corresponds to 79 dB SPL which is the level at which the speech stimuli is actually presented to human listeners in DAM testing. The results are referenced against subjective "SB" (Babbling) scores in our DAM database [2] by calculating the correlation coefficient $\rho_{subj,obj}$ between the predicted and DAM scores. The performance for female/male speech are slightly different, so $\rho_{subj,obj}$ is presented separately for female/male speakers as well as for overall (both male and female) performance.

Effect of misalignment

Our use of SFPs, like other intrusive methods of objective speech quality measurement involves the pre-alignment of the reference and degraded speech. This experiment attempts to investigate the effect of introducing an artificial delay between the original and degraded speech by shifting the degraded speech forward by Δ_t time. Other components of the algorithm are left intact for the subsequent computation of the correlation coefficient $\rho_{subj,obj}$ (between DAM "SB" score and the predicted value) as a function of Δ_t. The results are shown in Figure 7. The correlation coefficient $\rho_{subj,obj}$ are fairly invariant until the introduced delay reaches $1.2ms$, after which it starts decreasing. However, even with a delay of $6ms$, the $\rho_{subj,obj}$ only goes down by 0.05. This indicates that the TLD measurement algorithm using the SFPs is fairly tolerant to misalignment. The reason for the high tolerance is that the SFPs auto-align as described in the previous section, within a pitch period of the speech signals. Thus as long as the mis-alignment is within a pitch period, the performance is not expected to degrade significantly.

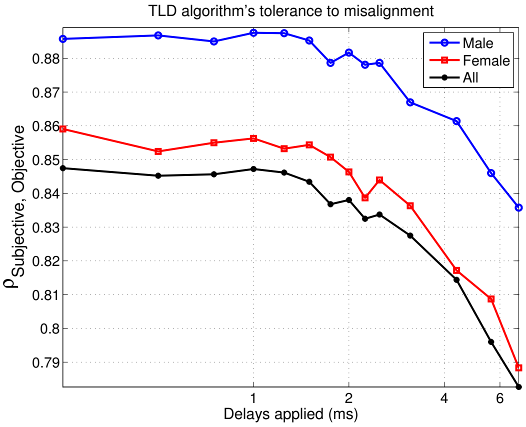

Figure 7: Performance of TLD objective measurement, as a function of artificially applied erroneous delay. When the delay is smaller than $1.2ms$, the accuracy $\rho_{subj,obj}$ is almost unaffected. Increasing the delay to larger than $1.2ms$ leads to lower accuracy at a fairly slow rate.

Effect of voiced section extraction

Typical examples of voiced sections identified by the algorithm are shown in Figure 8, where voiced section are marked as red. To measure the dependence of accurate identification of these segments, we will calculate the correlation coefficient after expanding the voiced sections arbitrarily to widths of $[t_{start} - \Delta_v, t_{end} + \Delta_v]$, where $[t_{start}, t_{end}]$ were the original conservative estimates of the width of the voiced regions. Correlation coefficients are then reported as a function of Δ_v. The tested in a range of Δ_v is $[-5, 60]ms$. Here, negative values indicate a shorter voiced section. The results are shown in Figure 9. The prediction accuracy $\rho_{subj,obj}$ reaches a maximum with Δ_v equal $18ms$ (for all speech) indicating the use of an overly conservative estimate of the endpoints of the voiced sections, previously. However, further expansion of the segments degrade the accuracy of prediction, indicating that the original hypothesis of constraining the analysis to voiced regions is quite valid.

Effect of different active speech level

According to [5], speech signals should be normalized to $-26dBoV$ before being played back to subjects. We undertake the same approach to ensure the exact conditions to simulate human perception as closely as possible. To simulate DAM testing, we

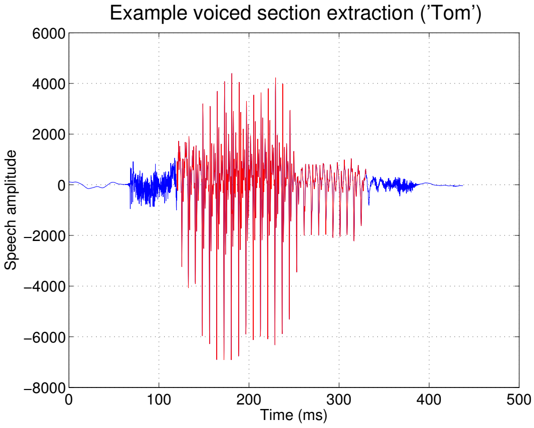

Figure 8: Example of identified voiced sections (in red). The algorithm is based on pitch estimates of the signal.

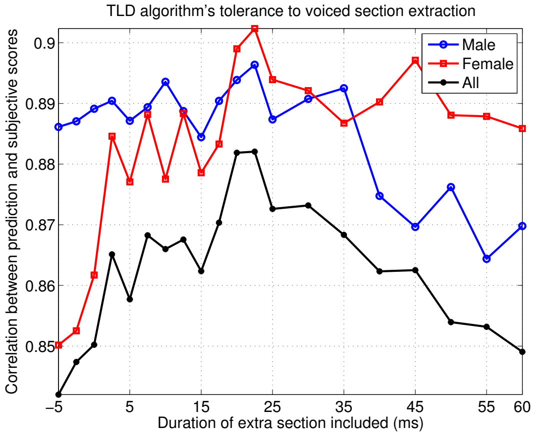

Figure 9: Performance of TLD objective measurement, as a function of extended lengths of voiced section extraction. The x-axis is Δ_v, which reflects an expansion of the voiced region to $[t_{start} - \Delta_v, t_{end} + \Delta_v]$.

further calibrate the cochlear model such that $-26dBoV$ represents 79 dB SPL. The cochlear-model is highly non-linear in nature - meaning that it is sensitive to the level at which stimuli is introduced. This experiment will try to examine our TLD measurement accuracy as a function of erroneous speech level, L. The original speech will continue to be normalized to $-26dB$, while the degraded speech will be normalized to different levels, ranging from $-21dB$ to $-51dB$. The results are shown in Figure 10. The correlation coefficient $\rho_{subj,obj}$ reaches a maximum when the degraded signal is scaled to $-26dB$ for male speech and $-31dB$ for female speech. When the level is decreased, $\rho_{subj,obj}$ decreases slightly. This high tolerance (or relative insensitivity) to level is attributed to the fact that the TLD prediction algorithm is based on the jitter of $SPF_{dis(t)}$. Thus even if the stimuli is scaled differently, any fluctuation of the temporal characteristics remain intact. This ensures that the active speech level does not change the accuracy of TLD measurement.

The characteristic is useful for future extension of the algorithm to non-intrusive and/or real time application to assess TLD distortions. In such circumstances, the original speech may not accessible. However, the SFPs themselves contains enough information to predict TLD, as long as current input speech is within a certain acceptable level.

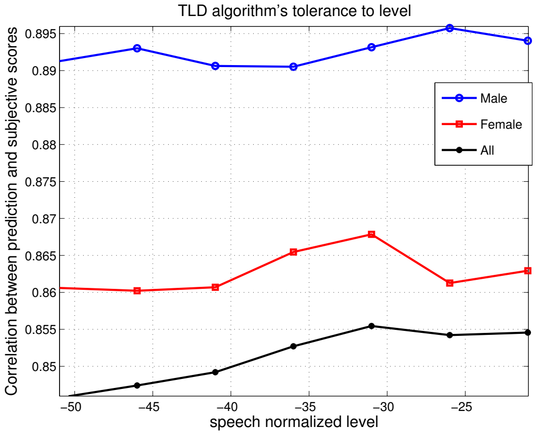

Figure 10: Performance of TLD objective measurement, as a function of active speech levels.

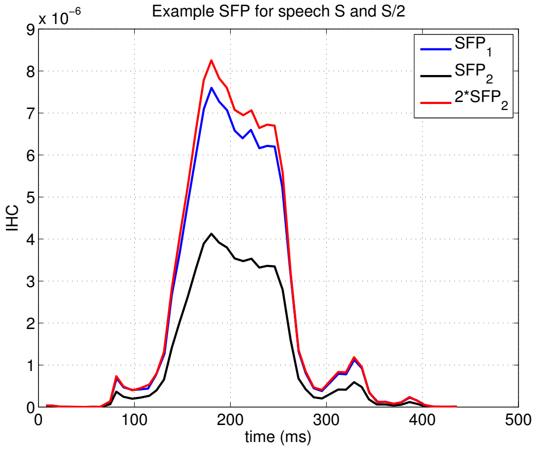

Figure 11: Non-linearity of CM output. SFP_1 and SPF_2 are calculated from Cochlear Model output of speech scaled by S and $S/2$ respectively. The two sets of SFPs are not identical. However, the fluctuation structure is preserved, which ensures the accurate prediction of TLDs.

DISCUSSION

This paper analyses the tolerance of an algorithm that predicts the perception of temporally localised distortions by artificially introducing errors in several parameters. We have shown that the algorithm has a high tolerance for misalignement between the original and distorted signal. This means that even if delays between the original and degraded speech are not totally eliminated, accurate results can still be achieved. This can be extremely useful when the signals have variable delays - as is often the case in VoIP networks. Additionally, the algorithm is also fairly invariant to speech level normalization. Introducing the reference and distorted signals at different levels have not shown significant impact on accuracy. This is because the SFPs preserve TLD information irrespective of level differences.
Smaller active speech levels do degrade the accuracy slightly. The investigation on selection of voiced sections reveal that the previously reported voiced/unvoiced selection can be optimized by expanding the segments by $18ms$. However, the results also shows that the hypothesis that quality evaluation is best achieved by restricting analysis to the voiced sections is well founded.

REFERENCES

[1] JG Beerends and JA Stemerdink. A perceptual speech-quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 42(3):115–123, 1994.

[2] Dynastat, INC. Diagnostic acceptability measure (DAM): A method for measuring the acceptability of speech over communication systems. Specification DAM-IIC, Dynastat. 1995.

[3] Donald D. Greenwood. A cochlear frequency-position function for several species – 29 years later. *Journal of the Acoustical Society of America (JASA)*, 87(6):2592–2605, 1990.

[4] Joseph L. Hall. Application of multidimensional scaling to subjective evaluation of coded speech. *JASA*, 110(4):2167–2182, 2001.

[5] ITU-T Recommendation P.56. Objective Measurement of Active Speech Level. *ITU-T*, 1993.

[6] ITU-T Recommendation P.862. Perceptual evaluation of speech quality(pesq), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *ITU-T*, 2002.

[7] Wenliang Lu and D. Sen. Extraction and tracking of formant response jitter in the cochlea for objective prediction of SB/SF dam attributes. *INTERSPEECH*, 2008.

[8] Wenliang Lu and D. Sen. Analysis of salient feature jitter in the cochlea for objective prediction of temporally localized distortion in synthesized speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[9] Wenliang Lu and D. Sen. Extraction of cochlear processed formants for prediction of temporally localized distortions in synthesized speech. *ICASSP*, 2009.

[10] S.R Quackenbush, T.P Barnwell III, and M.A Clements. *Objective Measurement of Speech Quality*. Prentice Hall, 1988.

[11] D. Sen. Predicting foreground SH, SL and BNH DAM scores for multidimensional objective measure of speech quality. *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 1:I–493–6, 17-21 May 2004.

[12] D. Sen. Determining the dimensions of speech quality form PCA and MDS analysis of the diagnostic acceptability measure. *MESAQIN*, 2001.

[13] D. Sen and Wenliang Lu. A framework for predicting speech quality using detectability of multiple distortions. *Proc. 38th International Audio Engineering Conference*, 38, June, 2010.

[14] W. D. Voiers. Diagnostic acceptability measure for speech communication systems. *Proc. IEEE ICASSP*, 1977.