Extraction and tracking of formant response jitter in the cochlea for objective prediction of SB/SF DAM attributes

Wenliang Lu, D. Sen

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

wenliang.lu@student.unsw.edu.au, dsen@ee.unsw.edu.au

Abstract

In this paper, we focus on the objective prediction of two of the foreground perceptual quality elements of the Diagnostic Acceptability Measure (DAM) - that of SB and SF - and show that they are correlated with statistical characteristics of features extracted from a physiologically motivated cochlear model response. The work complements earlier work where two other DAM quality elements, SH and SL, were predicted using the same cochlear model [1]. Novel methods of extracting salient features from the cochlear response as well as tracking their evolution are described. Finally, it is shown that the standard deviation of the features is highly correlated with the perception of "fluttering" (SF) and "babble" (SB) like distortions.

Index Terms: Objective measurement of speech quality, Diagnostic Acceptability Measure.

1. Introduction

The deployment of a multitude of speech coding and synthesis systems on telecommunication networks as well as in auditory prosthetic systems makes the accurate evaluation and monitoring of speech quality an important field of research. Despite significant gains in the field of objective measurement, the most accurate/reliable method of evaluation remains subjective testing. Typical subjective evaluation methods include the Mean Opinion Score (MOS) and the Diagnostic Acceptability Measure (DAM) [2]. While MOS testing provides an unidimensional quality score to any given speech system, the DAM evaluates the quality on a multidimensional distortion axes - ranging from "interrupted" to "tinny".

The ITU standardised objective measure - Perceptual Evaluation of Speech Quality (PESQ) [3] (ITU-T recommendation P.862 and associated addendums) - is inappropriate, according to the standard, for evaluating low bit-rate vocoders (below 4kbps) [3] as well as speech degraded by environmental conditions such as babble and military vehicle noise. 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 [4] - mainly due 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.

The work described in this paper is based on the premise that the remaining inadequacies of PESQ can be resolved - resulting in higher accuracy objective measures of speech quality - when explicit neuro-physiological models of audition are used in the place of PMMs. Further, in the same vein as DAM, and in line with our previous research [5], we consider the speech quality space to be multi-dimensional. As such we hypothesize that the objective prediction of the individual orthogonal dimensions of the quality space will lead to further increased accuracy. An added benefit of this approach is the ability to discern the type of distortion - something completely lost with the use of the unidimensional MOS measure or PESQ. In a previous paper, it was shown using Principal Component Analysis performed on a database of DAM scores, that the perception of speech quality can be described using three orthogonal dimensions [5]. The three dimensions are, temporally localised distortions, frequency localised distortions and those that are neither entirely localised in time or frequency. The temporal distortion dimension was found to be composed of the SI, SD, SB and SF quality elements of DAM. Of these, SB and SF are highly correlated to each other (see Fig. 1 in [5]). Given the subsequent success of predicting frequency localised distortions SL and SH [1], the focus of the current paper is an attempt at predicting SF and SF - from the family of temporally localised distortion elements.

2. Cochlear Response Feature Extraction

2.1. Cochlear Model

As mentioned in last section, the performance of PESQ can be largely attributed to the use of a PMM. The PMM however, is a very approximate estimation of the Basilar Membrane (BM) response. As such, it is not able to predict a number of linear and non linear characteristics of the true physiological response of the cochlea [6] - and corresponding psychophysics. An explicit physiological model of the cochlea, on the other hand, is not burdened by the drawbacks of PMM and is able to provide extremely precise details about how the cochlear behaves in response to auditory stimuli. The cochlear model (CM) used in this paper is a two-dimensional hydro-mechanical model [6, 1], which computes various electrical and mechanical responses in the cochlea. In particular, the model can be used to calculate BM and Inner Hair Cell (IHC) response as a function of time and space.

Our observation of the CM response is that it is highly redundant - due to the fact that the data is highly oversampled across the BM length. This necessitates dimensionality reduction and our strategy towards this has been to extract distinct features from the model response. In particular, we need to find features which correspond to the perception of the particular dimension of distortion (SB, SD, etc) that we are trying to predict.
2.2. Two Dimensional Evolution Tracking

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 Cochlear Model 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.

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 $0.0684\text{mm/sample}$ such that there are 512 discrete points across the approximate $3.5\text{cm}$ length of the human BM. It is possible to convert between place and frequency, using Greenwood’s map [7] (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 $l_p = [p_t, p_b]$ along the BM.

1. We start at the lowest boundary place $p_t$, which corresponds to the highest frequency in the region $[p_t, p_b]$. Find all local maxima along the time axis $CM_{p_t}(t)$, such that there are $M_{p_t}$ peaks at time $t_k; k = 1, 2, \ldots, M_{p_t}$. The peaks are chosen such that at time $t_k$, the cochlear response $CM_{p_t}(t_k)$ satisfies the criteria that it is larger than the $N$ neighbouring time sam-

![Figure 1: Cochlear response cross section 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.](image1)

![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.](image2)

![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.](image3)

![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.](image4)
bles, on either side of it, as follows: $CM_{p} (t_{k}) > CM_{p} (t_{k} - 1) > CM_{p} (t_{k} - 2) \cdots > CM_{p} (t_{k} - N)$, and $CM_{p} (t_{k}) > CM_{p} (t_{k} + 1) > CM_{p} (t_{k} + 2) \cdots > CM_{p} (t_{k} + N)$. 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 $(p_{c}, p_{b})$. The position of the peaks is stored in a matrix $PT$, such that $PT(p_{c}, k) = t_{k}$, $k = 1, 2, \cdots, M_{p}$. 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., $[t_{k,p-1} - t_{\text{backward}}, t_{k,p-1} + t_{\text{forward}}]$) 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, $t_{\text{backward}}$ 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.

2.3. Tracking Characteristic Place of Formants

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 2311Hz, 2420Hz 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 $\Delta t_{p} = \frac{1}{K} \sum_{k=2}^{K} (t_{p,k+1} - t_{p,k})$ (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 $\Delta t_{p}$, 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 $\Delta t_{p}$ for region 1 is lower than the $\Delta t_{p}$ 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 $1/T_{e}$ decreases with distance.

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 $\Delta t_{p}$, 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.

2.4. Center of mass for each formant region

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 the time period of the voiced /o/.
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.

3. Prediction of SB/SF from CM response

SB and SF are defined [8, 9] as "Babble" and "Fluttering" distortions respectively. From observation of systems which 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 [1], 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.

\[
jitter = \text{std}(\text{IHC}(SFP_{\text{ori}}) - \text{IHC}(SFP_{\text{dis}}))_{\text{voiced}} (1)
\]

In our tests, 9 different coding system were used, with 3 male and 3 female speakers. SFPs were extracted for voiced regions of the sentences and jitter was calculated as given by Equation 1. An important note here is that no DAM subjective scores were used or required during this process (for training). Plot of jitter versus SB is provided in Fig. 7. The resulting Pearson’s correlation coefficient for SB and SF were found to be −0.914 and −0.857 respectively.

4. Conclusion

The results above show that the process of extracting the SFP and the subsequent analysis of the IHC difference is highly correlated with human perception of SB and SF type distortions. This objective “jitter” distortion is a measure of the rate of difference in IHC response which was hypothesized to be closely associated with temporally localised distortions. The SFPs are closely linked to the formant or resonant frequencies of the vocal tract and represent the cochlear processed response in the time-place plane. The SFPs are however easier to locate and the use of an explicit physiological cochlear model nullifies the requirement for a Psychoacoustic Masking Model, albeit at the cost of computational complexity.

Future work will be focused on using the SFPs to predict SI and SD which are also triggered by temporally localised distortions. We predict that this will eventually produce a much more compact measure of speech quality than the multiple DAM quality elements which have been shown to be largely redundant [5].

5. References


