Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 865723, 11 pages
doi:10.1155/2009/865723

Research Article

Analysis of Salient Feature Jitter in the Cochlea for Objective Prediction of Temporally Localized Distortion in Synthesized Speech

Wenliang Lu and D. Sen

School of Electrical Engineering, University of New South Wales, Sydney, NSW 2052, Australia

Correspondence should be addressed to D. Sen, dsen@ee.unsw.edu.au

Received 13 October 2008; Revised 20 March 2009; Accepted 4 May 2009

Recommended by Jont Allen

Temporally localized distortions account for the highest variance in subjective evaluation of coded speech signals (Sen (2001) and Hall (2001)). The ability to discern and decompose perceptually relevant temporally localized coding noise from other types of distortions is both of theoretical importance as well as a valuable tool for deploying and designing speech synthesis systems. The work described within uses a physiologically motivated cochlear model to provide a tractable analysis of salient feature trajectories as processed by the cochlea. Subsequent statistical analysis shows simple relationships between the jitter of these trajectories and temporal attributes of the Diagnostic Acceptability Measure (DAM).

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) [1]. While MOS testing provides a unidimensional quality score to any given speech system, the DAM evaluates the quality on a multidimensional distortion axes—ranging from "interrupted" to "tinny".

The specification of the ITU-T recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ) [2], precludes its use for evaluation of low bit-rate vocoders (below 4 kbps) [2] 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 speech that has simply been distorted by low pass filtered speech ($f_c = 2$ kHz) as well as speech degraded by narrow band noise (from 400 Hz to 800 Hz). Even so, the PESQ algorithm betters earlier attempts at predicting MOS [3]—largely attributable to a highly evolved Psychoacoustic Auditory Model (PAM). The PAM is an attempt at modelling the linear component of the highly nonlinear hydromechanics of the human cochlea.

The work described within this paper is based on the premise that the inadequacies of PESQ can be resolved—resulting in higher accuracy objective measures of speech quality—when explicit neurophysiological models of audition are used in the place of PAMs. Further, in the same vein as DAM, and in line with previous research [4, 5], we consider the speech quality space to be multidimensional. As such, we hypothesize that the objective prediction of the individual orthogonal dimensions of the quality space will lead to further increase in 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 PCA performed on a database of DAM scores, that the perception of speech quality can be described using three orthogonal dimensions [4]. The three dimensions are, temporally localized distortions (PC1 in Figure 1), frequency

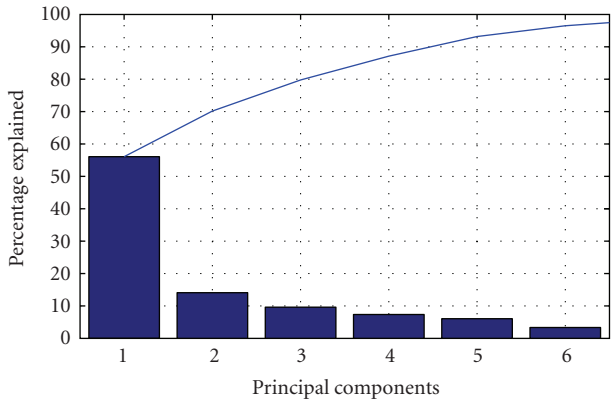

Figure 1: Principal component analysis of a database of DAM scores. PC1 comprises of temporally localized distortions consisting of SB, SF, SI, and SD, and accounts for 55% of the variance in overall quality. PC2 comprises mostly of frequency localized distortions which consists of SH and SL and accounts for 15% of the variance. PC3 includes SN and ST and accounts for less than 10%. Note that the first two components alone account for 70% of the variance [4].

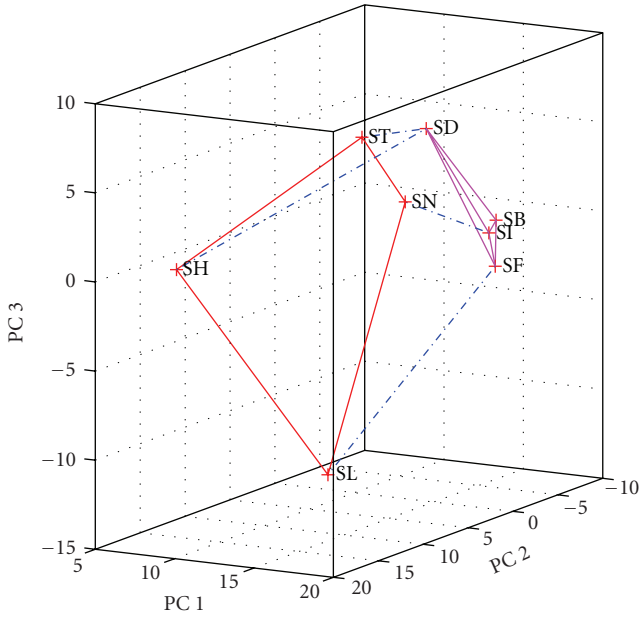

Figure 2: PCA analysis of DAM foreground attributes [4]. SH and SL stay at one side, while the temporally localized distortions are on another side. However, SD is slightly deviated from other three, that is, SB, SF, and SI.

localized distortions (PC2 in Figure 1), and those that are neither entirely localized 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, SF and SI are highly correlated with each other, as illustrated in Figure 2. The frequency localized distortions SL and SH were successfully predicted in earlier work [6]. The focus of the current paper is an attempt at predicting the family of temporally localized distortions, which account for 55% of the total variance in overall quality. The frequency localized distortions, in comparison, contribute 15% of the total variance.

In this paper, we propose a new methodology to extract features from a cochlear model response, to predict the perceptibility of temporally localized distortions. The paper is organized as follows. Section 2 discusses the cochlear model and explains the feature extraction process. Section 3 discusses the prediction of temporally localized distortions using the extracted features, followed by experimental results, and a discussion of the overall methodology.

2. Cochlear Response Feature Extraction

2.1. The Cochlear Model and the Motivation for Its Use. The performance of PESQ can be largely attributed to the use of a PAM. The PAM, however, is a functional model that approximates simultaneous masking. It can be treated only as an approximate estimation of the Basilar Membrane (BM) response. A primary failing of the PAM in the context of the current work of isolating and distinguishing temporal distortions is its lack of temporal resolution. To achieve high temporal resolution, the analysis frames used in PAM would be required to have compact time support. This would however render inadequate frequency resolution—inherently necessary for the PAM to produce accurate results. Moreover, the spreading functions (or filters) typically used in PAM to model BM functionality reduce frequency resolution, which is not reflective of the BM travelling waves. These carry far more spatiotemporal detail than can be observed by the PAM. It may be argued that such detail is not necessary to predict human perception. However, it is also true that not all of the loss in resolution depicted by the use of PAM is due to cochlear mechanics. Some of it is likely to be at higher stages of the auditory neurophysiology pathway. The methodology adopted in this work involves using the cochlear response resolution to first identify salient features and, only when the features have been detected and tracked, to reduce the resolution to a level that is representative of human perception. This strategy is impossible if a PAM is to be used to reduce time-frequency resolution as a front-end acoustic model.

Further, the linear characteristic of PAM means that it is not able to predict a number of nonlinear characteristics of the true physiological response of the cochlea [7], such as two-tone-suppression and cochlear emissions, and corresponding psychophysical phenomenon such as Upward Spread of Masking (USM) and loudness [8]. An explicit physiological model of the cochlea, on the other hand, is not burdened by the drawbacks of a PAM and is able to provide precise and high resolution spatiotemporal response of the cochlea due to auditory stimuli. In a latter section of this paper we discuss and compare the characteristics of the PAM as well as spectrograms with cochlear model output in the context of the current work. In particular we show that the PAM output lacks the resolution to carry out the analysis described within this paper, and that a consistent gain in prediction is achieved from using a nonlinear cochlear model (when compared with a linear cochlear model).

The cochlear model (CM) used in this paper is a spatially 2D hydromechanical model, 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. A block diagram of the cochlear model depicting the transduction path from the acoustic stimuli to its eventual transduction is shown in Figure 3. While detailed aspects of the cochlear model are beyond the scope of this paper, they may be found in various publications [6, 7, 9, 10]. The cochlear model can be broadly divided into three components: the *macromechanical model*, the *micromechanical model*, and the *nonlinear elements*. The ear-canal and ossicles are modelled as a linear filter—shown simply as "Middle Ear" in Figure 3. Various benchmarks comparing the model output to physiological and psychophysical data have been carried out to verify the performance of the model [7–9].

The macromechanical model is concerned with the dynamics of the fluid filled scalae and the Organ of Corti along the length of the cochlea. Of particular relevance is the travelling wave type mechanical response of the basilar membrane (BM). A Green's function [10] is used to numerically solve (in the time domain) the differential equations that result from assumptions of continuity (or conservation of fluid mass), inviscid and incompressible cochlear fluid loaded by the mass/stiffness and damping of the fluid and structures along the length of the cochlea. Spatial sampling is achieved by linearly discretizing the cochlea at 512 points along the 3.5 cm length of the cochlea.

The micromechanical model is concerned with the cilia (submerged between the tectorial membrane and the BM) and the associated Inner (IHC) and Outer (OHC) Hair Cells. The movement of the cilia are modelled as the direct result of the shear force created within the subtectorial space as a result of the relative movement of the BM to the tectorial membrane (TM). The TM is modelled as a transmission line, terminated by the cilia [9]. The phenomenological result of the micromechanical model is a cilia response that reflects an attenuated BM response basal to the Characteristic Place (CP). The cilia displacements are rectified and low-passed to derive the OHC and IHC receptor potentials. The IHC and OHC models are thus alike except for a high-pass filter that precedes the IHC model to account for the fact that the IHC cilia are not attached to the TM, but are driven by viscous fluid drag [11]. The IHC response from the model are reflective of receptor potentials, however no attempt is made to normalize them to units of Volts. Throughout the paper, it is these IHC responses that have been used as the output of the cochlear model and referenced as the CM response.

Cochlear nonlinearity imposed by OHC motility is modelled as mechanical feedback from the OHC, which modifies the macromechanical impedance. This is shown in Figure 3 as "Mechanical feedback." This is a cycle-by-cycle effect meaning an almost instantaneous feedback path in the model. The second and slower feedback due to efferent nerve fibres is not modelled within the model.

The model is implemented completely in the time domain. Due to discretization methods used in the model, as well as noise considerations inherent in nonlinear feedback systems, stability of the model is guaranteed only when it is run at a sampling rate considerably above the Nyquist sampling rate [10]. To adhere to this requirement, the 8 kHz sampled acoustic stimuli used in this work were upsampled by a factor of six before being processed by the cochlear model. Input to the cochlear model is on a sample by sample basis. Thus, for every sample into the model there is effectively a frame of 512 points of spatial data at the output. We discard every five of six frames, which has the effect of temporal downsampling back to 8 kHz.

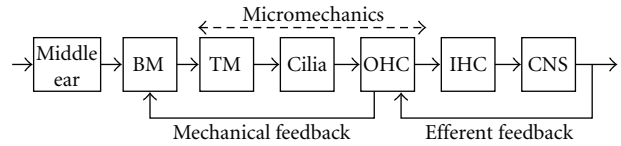

Figure 3: A computational model of the cochlea, used in this paper, showing the transduction path in auditory physiology.

A drawback of the use of the CM model is that it is highly redundant—due to the fact that the output is a 512 times oversampled relative to the input stimuli. This necessitates dimensionality reduction and our strategy towards this has been to extract distinct features from the model response. In particular, we isolate features which correspond to the perception of the temporally localized distortions—the focus of this paper.

2.2. 2D Evolution Tracking. The 2D cochlear Model response across time $\mathrm{CM}_p(t)$, at a single discrete place p (of arbitrary units), is a quasiperiodic 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_f can also be observed on the smooth low-passed envelope of the signal $e_p(t) = E\{\mathrm{CM}_p(t)\}$. This periodicity is due to the pitch of the speaker and is independent of place p (except for a slow evolution across space). These T_c, T_f are shown for a typical voiced section in Figure 4.

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. Exactly the 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. A peak tracking algorithm is used in Figure 5 to illustrate this evolution for a voiced section of speech.

We hypothesize that these peak tracks of the cochlear Model response are 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 Figure 6.

The output of the cochlear model is 2D data across time and space. The spatial sampling is 0.0684 mm/sample, such

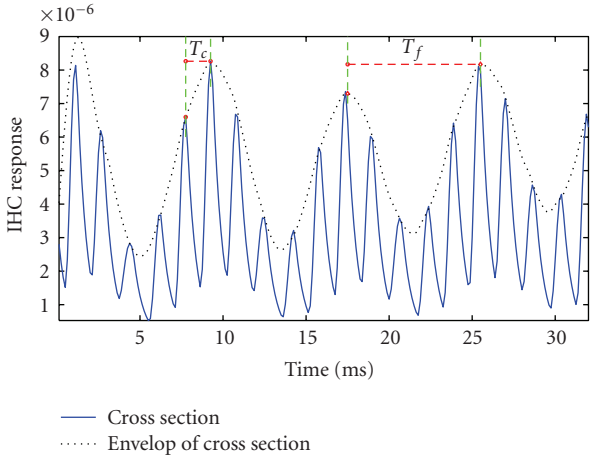

Figure 4: Cochlear response cross-section for voiced speech. Two types of periodicity, T_c and T_f, can be observed. T_c is given by the characteristic frequency of the place where the cross-section is taken, while T_f is determined by the fundamental frequency of this speech segment.

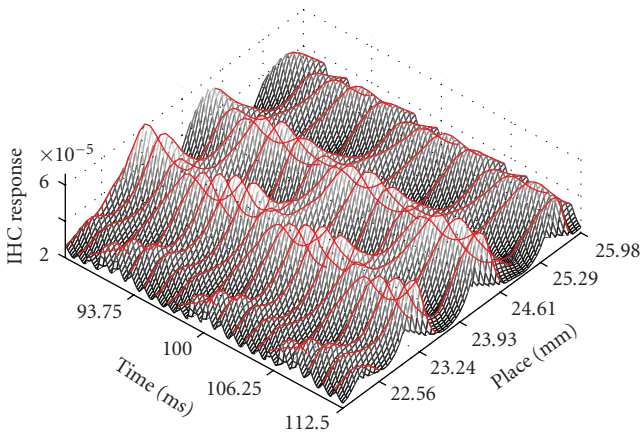

Figure 5: Cochlear response as a function of time and place, with peak tracks for a voiced segment of speech (/o/). Dark lines indicate the peaks or crests of the response, and exhibit a regular, quasiperiodic structure which is also evidenced in Figure 4.

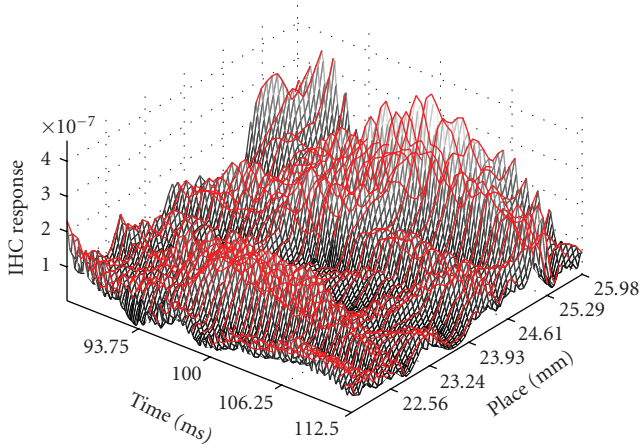

Figure 6: Peak tracks from the cochlear response for an unvoiced segment of speech (/s/). The quasiperiodic structure that appears in Figure 5 is not present.

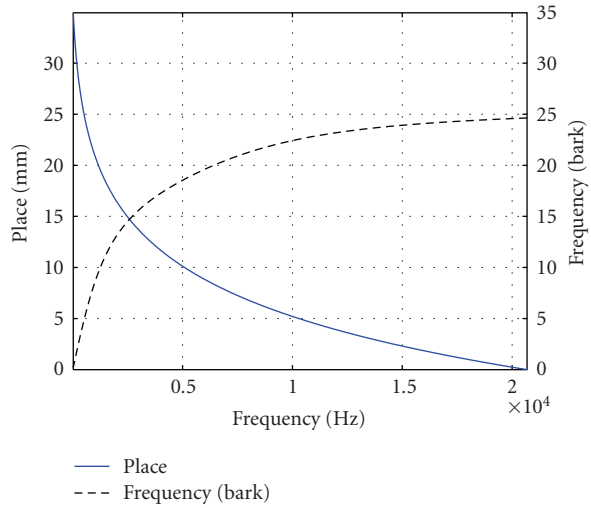

Figure 7: Relationship between frequency (in Hz and Bark) and place. Place can be converted to frequency using Greenwood's map [12].

that there are 512 discrete points across the approximate 3.5 cm length of the human BM. The relationship between place and frequency can be approximated using Greenwood's map [12]. This mapping is however only valid at threshold levels. To provide an indication to the reader, 24 mm along the cochlear length represents the characteristic place for a 600 Hz sinusoid (at threshold), as can be seen in Figure 7.

The steps below describe an algorithm to track the 2D evolution of the cochlear response $\mathrm{CM}_p(t)$ on a closed spatial region $p = [p_l, p_h]$ along the BM, where p_l and p_h are the lower and upper bounds along the place axis with $p_l, p_h \in [1, 512]$.

(1) We start at the lowest boundary place p_l, which corresponds to the highest frequency in the region $[p_l, p_h]$. All local maxima along the time axis $\mathrm{CM}_{p=p_l}(t)$ are found, 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 $\mathrm{CM}_{p_l}(t_k)$ satisfies the criterion that it is larger than the N neighbouring time samples, on either side of it, as follows: $\mathrm{CM}_{p_l}(t_k) > \mathrm{CM}_{p_l}(t_k - 1) > \mathrm{CM}_{p_l}(t_k - 2) \cdots > \mathrm{CM}_{p_l}(t_k - N$, and $\mathrm{CM}_{p_l}(t_k) > \mathrm{CM}_{p_l}(t_k + 1) > \mathrm{CM}_{p_l}(t_k + 2) \cdots > \mathrm{CM}_{p_l}(t_k + N)$. The value of N is a function of the temporal sampling rate and is empirically determined 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 $PT(p_c, k) = t_k, k = 1, 2, \ldots, M_{p_c}$. 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

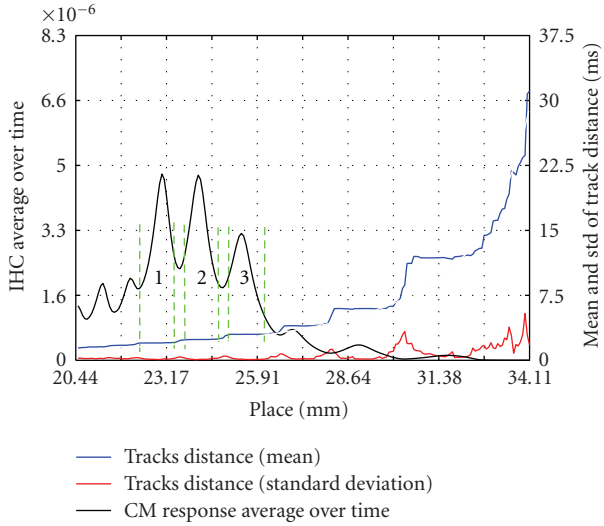

FIGURE 8: IHC response in comparison to track distances. as function of place. The left vertical axis represents IHC response while the right vertical axis represents time difference between the tracks (in milliseconds). The line in blue, red, and black represents mean track distance, standard deviation of the track distances and the average IHC response over time. It may be observed that the track distance rises in discrete steps, with small steps corresponding to high IHC response, and high steps corresponding to low IHC response and high standard deviation. Three regions labelled as "1", "2", and "3" between dotted vertical lines, have been identified as Perceptually Relevant Regions (PRR). The plot corresponds to the same response in Figure 5 averaged over time and with an extended range in place.

neighborhood (i.e., $[t_{k,p-1} - t_{\text{backword}}, t_{k,p-1} + t_{\text{forward}}]$) of each peak position from the previous place, $p-1$. Due to causality, the peak tracks always move towards increasing time and place. For this reason, t_{backward} can be small. 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 the track is terminated at place $p-1$ and no further search along this track is performed in the future. 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 postprocessing involves connecting broken tracks which are possibly part of the same track, and 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.

An example of the above steps is illustrated in Figures 5 and 6. The continuous lines capture information related to 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 space) and thus result in almost parallel looking tracks.

2.3. Locating Perceptually Relevant Regions. Articulatory features such as vocal tract resonances (formants) and pitch harmonics are easily distinguishable in the 2D rendering of the CM response. During voiced speech, these features are distinguishable as distinct "peaks" or high energy regions in the CM response, as can be observed in Figure 5. In the figure, three pitch harmonics at the first formant region can clearly be tracked over time and place. They appear at approximately 23.11 mm, 24.20 mm, and 25.57 mm from the base of the BM, their positions changing slightly with time. These places correspond to approximately 710 Hz, 590 Hz, and 463 Hz. Instead of referring to these in terms of articulatory features, it is more appropriate to refer to these as Perceptually Relevant Regions (PRR), 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 PRRs extremely accurately over time and place. What is actually being tracked is the effect of the articulatory features as processed by the cochlea. This is one of the main reasons that the use of CM response is far superior to the use of a spectrogram or a PAM, as the CM response reflects only the information that remains after nonlinear cochlear processing.

One of the important features of the PRRs 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 PRRs. This is demonstrated in Figure 4.

The next step in our feature extraction is to focus on just the PRRs. This is facilitated by the observation that the average time difference between the peak tracks $\overline{\Delta_{t_p}} = (1/(K-1)) \sum_{k=1}^{K-1} (t_{p,k+1} - t_{p,k})$ (over the duration of the voiced section) is almost constant across the region of each PRR, where k is the index of track, and K is the total number of tracks. This is shown in Figure 8 which shows that in each of the three PRRs, 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 place decreases with increasing distance, reflecting the fact that the characteristic frequencies $1/T_c$ decrease with distance.

To focus on the PRRs, we use a two pronged strategy. First, we impose an energy threshold such that only sections of the CM response above the threshold are kept. In addition, we use the characteristic of the $\overline{\Delta_{t_p}}$, whereby it increases in (almost) discrete steps (as shown by the blue line in Figure 8). The boundaries of the plateaus further distinguish relevant regions. These regions are shown in Figure 9 as areas between horizontal lines ("PRR1", "PRR2", and "PRR3"). The three

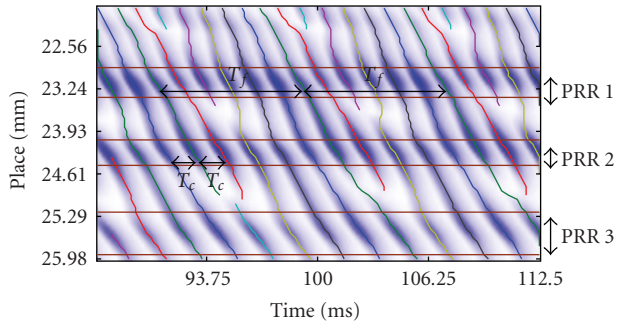

Figure 9: Cochlear response with peak tracks for voiced speech /o/ on the time-place plane. The parallel structure between tracks can be observed at the PRRs (between straight horizon lines). The three regions PRR1, PRR2, and PRR3 are the same three regions labelled in Figure 8 as "1", "2", and "3". Also, the T_c and T_f in Figure 4 are indicated here.

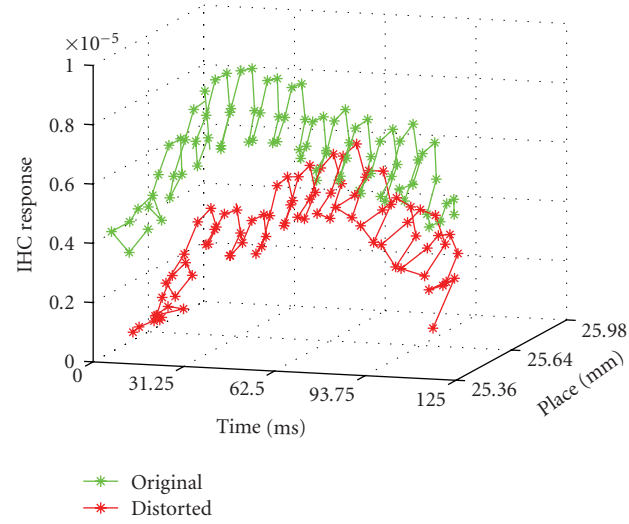

Figure 10: Center of Mass of tracks in one PRR. Notice the swirling characteristic of TCPs.

regions correspond to the three dominant pitch harmonics in the vicinity of the first formant.

2.4. Center of Mass for Each Formant Region. A characteristic of the peak tracks within each PRR is the fact that they are quasiparallel on the time-place plane (much more so than in other regions). To reduce the dimensionality and computational complexity, the "center of mass" of each track slot (restrained by PRRs) is computed. Each new point is characterised by a time, place, and amplitude, (τ, χ, R). We call these points Track Center Points (TCP). The amplitude is simply the average of the IHC responses constrained by the boundaries of a track. The time (τ) and place χ values are calculated using the following three equations:

$$R = \frac{1}{M}\sum_{i=1}^{M} \mathrm{IHC}_i,$$

$$\tau = \frac{\sum_{i=1}^{M} \mathrm{IHC}_i t_i}{\sum_{i=1}^{M} \mathrm{IHC}_i}, \tag{1}$$

$$\chi = \frac{\sum_{i=1}^{M} \mathrm{IHC}_i p_i}{\sum_{i=1}^{M} \mathrm{IHC}_i}.$$

Here IHC_i is the IHC amplitude, t_i is time position, and p_i is the place position, of point i. M is the number of points in one track. A typical set of consecutive TCPs (in one formant region) is plotted in Figure 10, which is inferred from PRR3 in Figure 9. The plot reveals a swirling 3D curve. The period of the swirl corresponds to the periodicity of the underlying (time domain) speech signal and is given by T_f, in Figure 4.

Corresponding TCPs across period T_f, are also similar in intensity and place—more so than neighbouring TCPs. In a further attempt at reducing dimensionality, each set of TCPs in a single period T_f is reduced to a single "center of mass" as given by (1). We call these points the Salient Formant Points (SFP), reflecting the fact that they are indicative of formant energy as a function of time and place. Time periodicity has been removed as a result of this final process. These corresponding SFPs between the original and distorted speech signals are highly synchronized in time. This is of great benefit as most intrusive objective speech quality measures, such as PESQ [2], require fairly complex preprocessing to synchronize the two signals accurately, a step for which our system can afford to be less precise due to this automatic SFP synchronization. Figure 11 indicates the final result of this process and shows the extracted Salient Formant Points (SFP) in 3D space of time, place, and IHC response. Figure 12 is a plot of the points showing the extraction times of the original and distorted signals, respectively. A most notable feature is that the points extracted in this manner for the two different systems are automatically synchronized, without an explicit requirement for the signals to be synchronized accurately at the input.

Figure 14 shows that the points are lightly dispersed over place due to the different coding systems, as should be expected. Finally, Figure 13 shows the IHC response at each of the extracted points. Note the significant amplitude difference between original and distorted signals. In our intrusive prediction for speech quality, original signals are used as a reference of "smoothness". A perceptual formant distance PFD is defined as below:

$$\mathrm{PFD} = |\mathrm{SFP}_{\mathrm{dis}} - \mathrm{SFP}_{\mathrm{ori}}|_{\mathrm{voiced}}. \tag{2}$$

The PFD is used to predict temporal distortions, as described in the next section. Note that in an extreme situation, if the original and distorted SFPs are parallel to each other in amplitude, the PFD is flat or constant, only reflecting a multiplicative constant between the two signals. It is the deviation along the time axis of the PFD that carries information on temporal distortions.

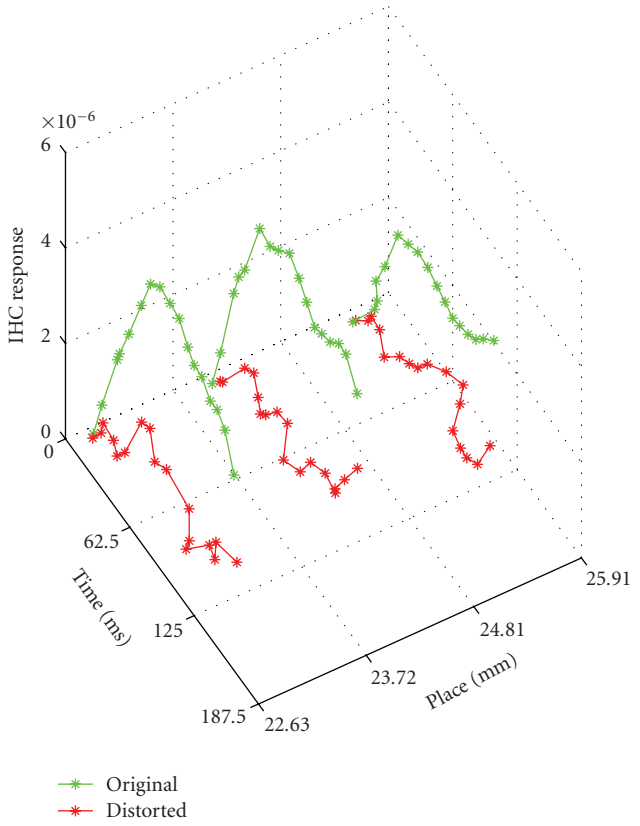

FIGURE 11: Extracted Salient Formant Points. Three sets of original and distorted (perceptual) formants are displayed. Both the original and distorted TCPs in Figure 10 are converted to SFPs.

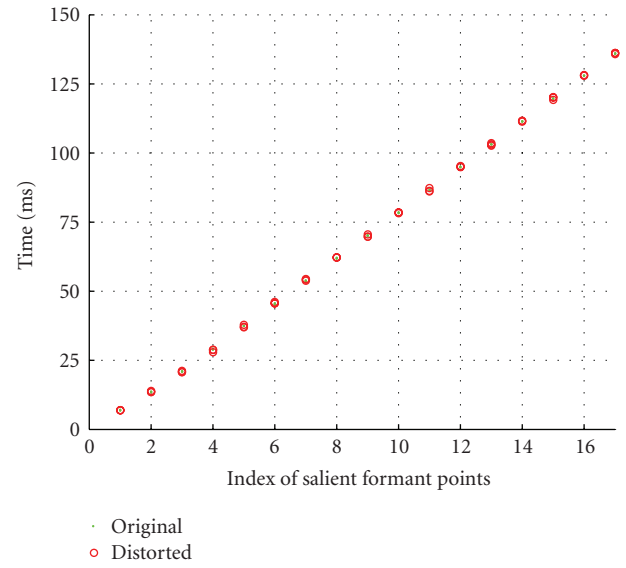

FIGURE 12: 2D projection of Figure 11 showing the time instances of the extracted SFP_{dis} and SFP_{ori}. Note that the time instances fall on top of each other—implying an automatic synchronization in time between the distorted and original signal.

3. Predicting Temporally Localized Noise

Unlike frequency localized distortions [6], temporally localized distortions are isolated over compact sections of the time axis. In contrast, frequency localized distortions extend over wide lengths of time, or indeed over the entire length of the signal (as would be the case for low-pass or high-pass speech). Temporally localized distortions have been represented using descriptors such as "clipping", "additive noise" and "fluttering" amongst others. In our observation, the temporally localized distortions can be further subclassified into a "rapid" and "slow" category depending on the rate at which the formants of the distorted signal vary with respect to the original signal. The "slow" category causes distortions that are typically described as "fluttering" and "babble" while the rapid category causes distortions that elicit "raspy" and "crackling" types of responses from listeners.

The above observation leads us to the hypothesis that temporally localized distortions are related to the rate at which the synthesized salient features deviate from the original in both time and frequency. A similar hypothesis relating "fluttering" distortions to "formant fluttering" was made in [13]. The PFD calculated in Section 2.1 combines the effect of formant deviations in cochlear response and place (frequency) and thus lends itself to the exploration of the above hypothesis. We estimate the rate of formant deviation (in the cochlea) (or "jitter") using the following two equations:

$$J_{\text{slow}} = K_1 \sum_{\text{voiced}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\text{PFD}_i - \overline{\text{PFD}}\right)^2}, \tag{3}$$

$$J_{\text{rapid}} = K_2 \sum_{\text{voiced}} \frac{\partial(\text{PFD})}{\partial t}. \tag{4}$$

Here K_1 and K_2 are constants. Equation (3) is well suited to the prediction of distortions in the slow category (of temporally localized distortions) while (4) is well suited for the prediction of distortions in the fast category. To test our hypothesis, we have attempted to predict the relevant attributes of a database of DAM subjective test scores. In particular, we have classified the SF, SI, and SB attributes of DAM to the second "slow" category of temporally localized distortions and the SD attribute of DAM to the "rapid" category.

The DAM specification [13, 14] defines SB, SF, and SI, as "Babbling", "Fluttering" and "Interrupted" distortion respectively. SD is defined as "Signal Rasping", and "Crackling" [13] and being caused by a broad range of factors, (e.g., center clipping, additive noise, etc.). One difference between SD and the other three, is that the former represents distortion that is localized over smaller lengths of time, implying rapid evolution of formants and eliciting a "harsh" perception amongst listeners.

The classification of these attributes to temporally localized distortions was based on earlier work [4], where it was shown that these attributes contribute almost 55% of

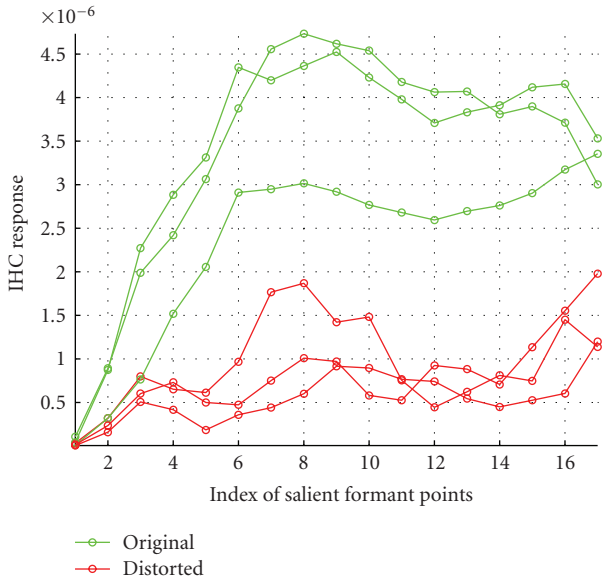

Figure 13: 2D projection of Figure 11 showing amplitude of the extracted SFP_{dis} and SFP_{ori}. The distance between original and distorted amplitude carries the temporally localized distortion information.

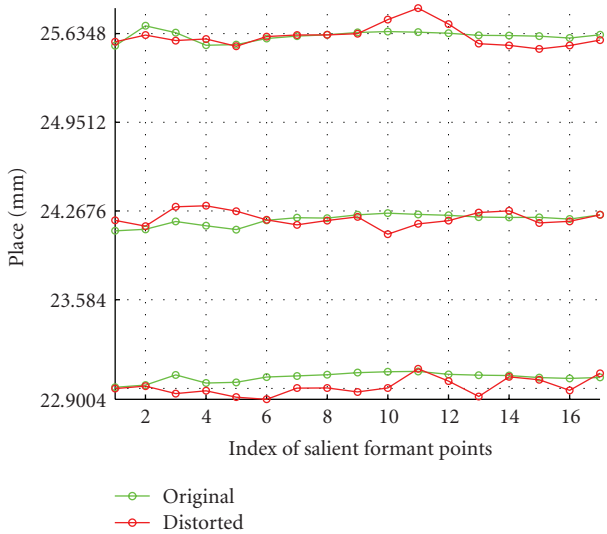

Figure 14: 2D projection of Figure 11 showing the (place) location of the extracted SFP_{dis} and SFP_{ori}.

the total variance in the subjective scores (as shown in Figure 1). It is interesting to note that while all four of these DAM attributes (SF, SI, SB, and SD) were classified as temporally localized distortion descriptors in [4], there was clear demarcation between SD and the rest as shown in Figure 2. In the next section we report on the results of using (3) and (4) to predict these DAM attributes.

4. Results

Nine different coding systems were tested, each with three male and three female speakers. The systems tested are shown in Table 1. There were thus a total of 54 candidates with different system and speaker combinations to be tested.

Table 1: Coding Systems represented in the database being under test.

Index	Codec name
1	original
2	G728_clean_0
3	LPspeech_clean_0, low pass filter, $f_c \approx 2$ kHz
8	G729_clean_0
11	emMELP52_quiet_0
14	MELPWI52_quiet_0
31	E1 (Combined Melp and WI, 5.2 kbps—in quiet)
32	F1 (G729a 8 kbps—in quiet)
49	E4 (MELPe_fix 1.2 kbps, 42 bit quantizer—1% random BER)
50	F4 (G729a 8 kbps—1% random BER)

For each candidate, we calculated an objective score in the "rapid" category, and another in the "slow" category as given by (4) and (3), respectively. We hypothesize that the "slow" score is correlated with all the three attributes of "SB", "SF", and "SI", due to their similarity shown in the PCA and MDS analyses while the "SD" attribute is correlated with the "rapid" category objective score.

The correlation coefficients ρ between the subjective DAM attributes [14] and corresponding predicted scores (from (3) and (4)) are calculated as follows:

$$\rho = \frac{\sum_{i=1}^{N}\left(S_i - \overline{S}\right)\left(O_i - \overline{O}\right)}{\sqrt{\sum_{i=1}^{N}\left(S_i - \overline{S}\right)^2 \sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2}}, \tag{5}$$

where S and O are the subjective (DAM) and objective (J_{slow} or J_{rapid}) prediction scores, respectively, and N is the number of candidates (54 in our case).

As hypothesized, the predicted score J_{slow} is highly correlated with all three temporal DAM attributes: SB, SF, and SI [15]. The correlation coefficients are $\rho_{SB,J_{slow}} = -0.91$, $\rho_{SF,J_{slow}} = -0.86$, $\rho_{SI,J_{slow}} = -0.81$. Figure 15 illustrates the relationship between the subjective SB scores and the objective prediction. Further improvements can be achieved by performing polynomial regression [13]. Our test results show that a second-order polynomial regression can improve the $\rho_{SB,J_{slow}}$ to 0.93.

SD, the only one attribute in the "rapid" category, is highly correlated with the prediction of J_{rapid}, which presents the correlation coefficient $\rho_{SD,J_{rapid}}$ of −0.89. Figure 16 reveals the relationship between SD subjective scores and objective predictions J_{rapid}. Like SB, the $\rho_{SD,J_{rapid}}$ can also be slightly improved to 0.90 with third-order polynomial regression.

5. Discussion

The results above show that the process of extracting and tracking (across space and time) salient features from a cochlear model output and their subsequent time rate of

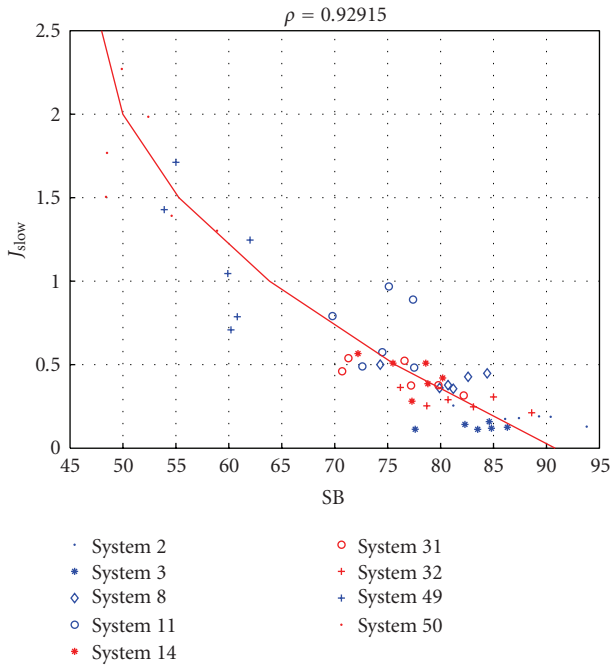

FIGURE 15: Scatter plot of SB versus J_{slow} of (3). The continuous line is a second-order line of best fit. The resulting correlation coefficient is $\rho = 0.93$.

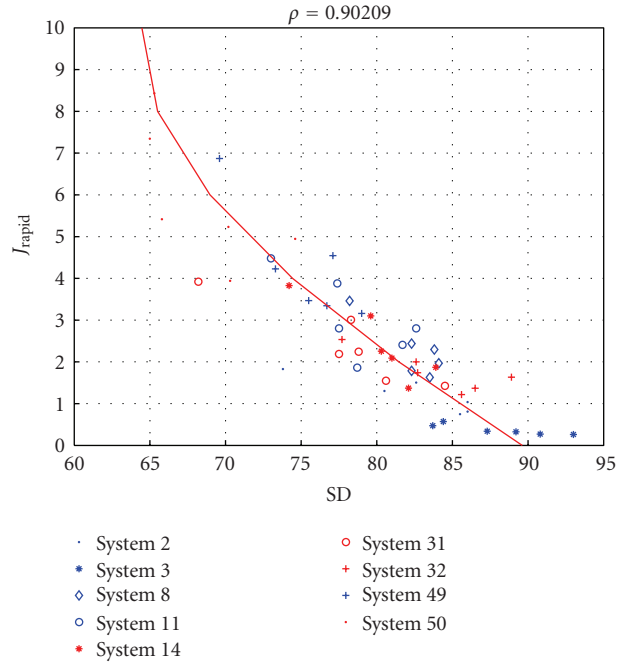

FIGURE 16: Scatter plot of SD versus J_{rapid} of (4). The continuous line is a third-order line of best fit. The resulting correlation coefficient is $\rho = 0.90$.

deviation in comparison to a feature set derived from a clean (undistorted) signal is correlated with the perceptibility of temporally localized distortions. The feature set that was extracted was broadly termed Salient Formant Points or SFPs. The SFPs are so named due to their association with the cochlear processed high energy formants and are clearly represented over the time-place dimension in the cochlear response.

The methodology described in the paper to extract temporally localised deviations is facilitated by the spatiotemporal resolution of the cochlear response. Figures 17, 18, and 19 show the output of the cochlear model, a psychoacoustic model (using a frame length of 1024 points) and a spectrogram (using a frame length of 1024 and an overlap such that one new sample was introduced at each frame). It is clear from these figures that the resolution afforded by the cochlear model is not available in either of the other two analysis methods. Indeed, when we blindly replace the CM with a PAM, the feature extraction/tracking algorithm was unable to perform as various characteristics of the response was just not present at the output of the PAM. The same is true if we were to replace the CM with a spectrogram. Increasing the temporal resolution of the PAM by taking shorter analysis frames renders it inaccurate in the frequency domain. Increasing the time resolution of the spectrogram does not produce an output, that is, reflective of the processing carried out by peripheral auditory processing.

One aspect of the cochlear model that makes it superior to simultaneous masking models (essentially the PAMs used in systems such as PESQ) is its ability to reproduce nonlinear phenomena. This is a direct result of incorporating the OHC

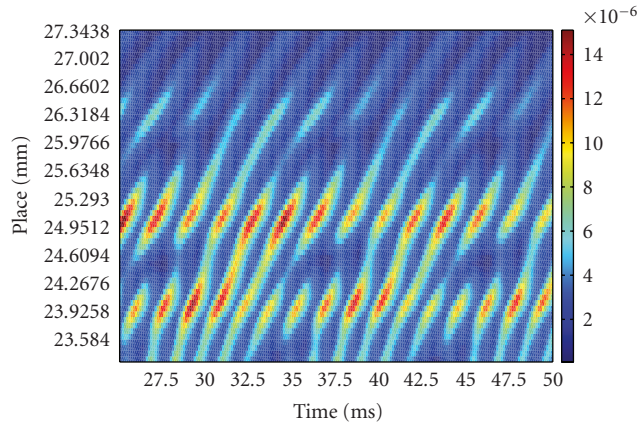

FIGURE 17: Cochlear Model response for */ai/* in the vicinity of the first formant. The y axis is place (mm), ranging from 23.2 mm to 27.3 mm, which correspond to approximately 694 Hz to 335 Hz. Figures 18 and 19 correspond to the same time segment of the speech signal.

mechanical feedback into the model. To test how much of an effect the nonlinearity has in predicting temporally localised distortions, we turned off the nonlinearity in the CM and ran an identical feature extraction, tracking and subsequent deviation analysis as described in this paper. The results are shown in Table 2 below. While the differences are not significantly high, the predicted results using a nonlinear model is higher than that using a linear model for three out of the four cases. A better test of course would be to use a subjective database where different loudness levels of

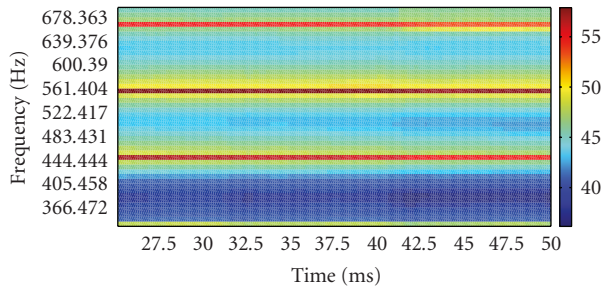

Figure 18: Response from a psychoacoustic model for response for */ai/*—the same segment of the speech signal as Figures 17 and 19. The *y* axis is frequency (Hz), ranging from 335 Hz to 694 Hz, which corresponds to 3.25 to 6.34 Bark frequency. The psychoacoustic model uses a frame length of 1024 and an overlap such that one new sample was introduced for each new frame.

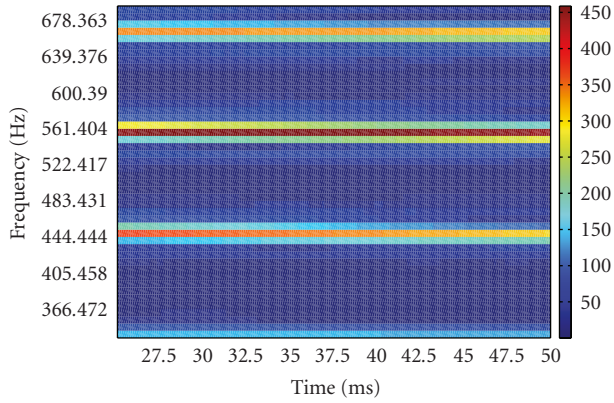

Figure 19: Spectrogram for first formant for */ai/*—the same segment of the speech signal as Figures 17 and 18. The spectrogram use a frame length of 1024 and an overlap such that one new sample was introduced at each frame.

Table 2: Correlation coefficients between subjective and predicted scores using linear and nonlinear cochlear models. The results for linear CM are consistently lower, compared to nonlinear CM, except for SF.

Distortions	Nonlinear CM	Linear CM
SB	−0.91	−0.88
SF	−0.86	−0.865
SI	−0.81	−0.78
SD	−0.89	−0.81

speech were tested. We did not have at our disposal such a database as the speech was always presented to the listeners at 79 dB SPL. The consistency in which the nonlinear model produces better prediction in our tests allows us to conjecture that when speech is presented at different levels, a nonlinear model of the cochlea will lend itself to more accurate predictions of distortion detectability.

The results of the current work match the PCA/MDS analysis carried out earlier. In the current work, DAM attributes of SB, SI, SF, and SD were empirically subclassified into two groups based on their rate of SFP evolution. "Fluttering" (SF), "babble" (SB), and "interrupted" (SI) types of distortions were observed to evolve at a slower rate than raspy (SD). This motivated the two proposed "jitter" distortion measures, J_{rapid} and J_{slow}. The former was used to predict SD, while the latter was used to predict SB, SF, and SI. The accuracy in prediction of these two classes of temporal distortion matched the earlier PCA/MDS analysis which showed high correlation between SB/SF/SI and the slightly differentiated SD.

Future work will be focused on the precise prediction of the Composite Acceptability Estimate (CAE) and MOS scores, both of which are unidimensional measurements of speech quality.

Definitions

MELP: Mixed excitation linear prediction
MELPe: Enhanced MELP
WI: Waveform interpolation
DAM: Diagnostic acceptability measure, one subjective speech quality developed by Dynastat Inc., USA. This set of measures put speech quality into a multidimensional space
SB: Babble, for example, systems with errors
SD: Harsh/raspy, for example, peak clipped speech
SF: Fluttering, for example, interrupted speech
SI: Interrupted, for example, packetized speech with clitches. SB, SF, SI, and SD are temporally localized distortions
SH: Thin, for example, high pass speech. Not like the above four distortions, SH and SL below are frequency localized
SL: Muffled, for example, low pass speech
CAE: Composite acceptability estimate. It present overall speech quality, based on other subjective parameters, for example, SB, SF, SH, etc
MOS: Mean opinion score
PESQ: Perceptual evaluation of speech quality, the current ITU-t standard for intrusive objective measurement of speech quality
CM: Cochlear model
PRR: Perceptual relevant region. Each region actually represent a perceptual pitch, while a few regions nearby group to be one perceptual formant
TCP: Track center point
SFP: Salient formant point. TCPs in one PRR are reduced to SFP for easier comparison between original and distorted systems
PCA: Principal component analysis
MDS: Multidimensional scaling.

Acknowledgments

The authors thank the two anonymous reviewers for their valuable suggestions, and Dynastat for providing us with a database of DAM scores.

References

[1] W. D. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '77)*, 1977.

[2] 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.

[3] J. G. Beerends and J. A. Stemerdink, "Perceptual speech-quality measure based on a psychoacoustic sound representation," *Journal of the Audio Engineering Society*, vol. 42, no. 3, pp. 115–123, 1994.

[4] D. Sen, "Determining the dimensions of speech quality form PCA and MDS analysis of the diagnostic acceptability measure," in *Proceedings of the International Conference on Measurement of Speech and Audio Quality in Networks (MESAQIN '01)*, 2001.

[5] J. L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *Journal of the Acoustical Society of America*, vol. 110, no. 4, pp. 2167–2182, 2001.

[6] D. Sen, "Predicting foreground SH, SL and BNH DAM scores for multidimensional objective measure of speech quality," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 1, pp. I493–I496, May 2004.

[7] D. Sen and J. Allen, "Benchmarking a two-dimensional cochlear model against experimental auditory data," in *Proceedings of the MidWinter Meeting on Association for Researchin Otolaryngology (ARO '01)*, February 2001.

[8] J. Allen and D. Sen, "A unified theory of two-tone suppression and upward-spread of masking," *The Journal of the Acoustical Society of America*, vol. 103, p. 2812, 1998.

[9] D. Sen and J. B. Allen, "Functionality of cochlear micromechanics—as elucidated by upward spread of masking and two tone suppression," *Acoustics Australia*, vol. 34, no. 1, pp. 37–42, 2006.

[10] J. B. Allen and M. M. Sondhi, "Cochlear macromechanics: time domain solutions," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 123–132, 1979.

[11] P. Dallos, "Response characteristics of mammalian cochlear hair cells," *Journal of Neuroscience*, vol. 5, no. 6, pp. 1591–1608, 1985.

[12] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, 1990.

[13] S. Quackenbush, T. Barnwell III, and M. Clements, in *Objective Measurement of Speech Quality*, A. V. Oppenheim, Ed., Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

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

[15] W. Lu and D. Sen, "Extraction and tracking of formant response jitter in the cochlea for objective prediction of SB/SF dam attributes," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '08)*, Brisbane, Australia, September 2008.