EXTRACTION OF COCHLEAR PROCESSED FORMANTS FOR PREDICTION OF TEMPORALLY LOCALIZED DISTORTIONS IN SYNTHESIZED SPEECH

Wenliang Lu and D. Sen

University of New South Wales, Sydney, Australia

ABSTRACT

Temporally localized distortions account for the most variance in subjective evaluation of coded speech signals [1, 2]. 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 trackable analysis of formant 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).

***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) [3]. 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) [4] (ITU-T recommendation P.862 and associated addenda) - is inappropriate, according to the standard, for evaluating low bit-rate vocoders (below 4kbps) [4] 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 [5] - 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 [1], 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 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. It was shown using Principal Component Analysis performed on a database of DAM scores, that the speech quality can be described using three orthogonal dimensions [1]. The three dimensions are, temporally localized distortions (*PC1* in Fig 1), frequency localized distortions (*PC2* in Fig 1) and those that are neither entirely localized in time or frequency. The frequency localized distortions, i.e., SL and SH, were successfully predicted in earlier work [6]. The frequency localized distortions contribute 15% of variance to overall quality, while the largest component, temporally localized distortions, take up to 55%. The focus of the current paper is an attempt at predicting the family of temporally localized distortion elements, which was found to be composed of the SI, SD, SB and SF quality elements of DAM.

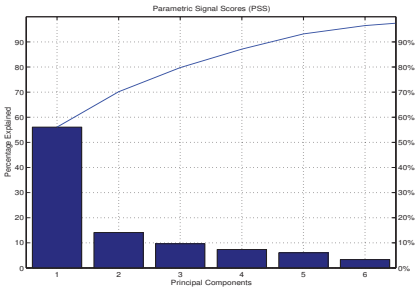

Fig. 1. PC1 is temporally localized distortions, which is consist of SB/SF/SI/SD, and account for 55% percentage of variance to overall quality. PC2 is frequency localized distortions, which contains SH/SL. PC3 includes SN/ST. Note that the first two components add up to be 70%.

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 describe a number of linear and non linear characteristics of the true physiological response of the cochlear [7] - and corresponding psychophysics. An explicit physiological model of the cochlea, however, 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, 7], which computes various electrical

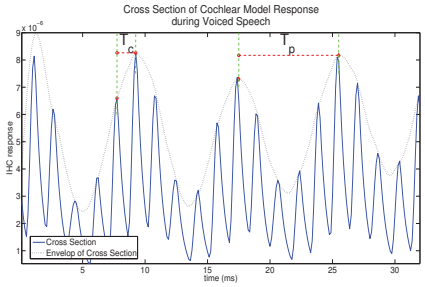

Fig. 2. *Cochlear response cross section for voiced speech. Two types of periodicity, T_c and T_p, can be observed. T_c is given by the characteristic frequency of the place where the cross section is taken, while T_p is determined by fundamental frequency of this speech segment.*

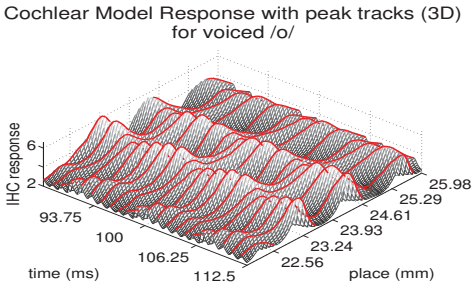

Fig. 3. *Cochlear response as a function of time and place, with peak tracks for an voiced segment of speech (/o/). Dark lines indicate the peaks or crests of the response, and exhibit a regular, quasi-periodic structure which is also evidenced in Fig. 2.*

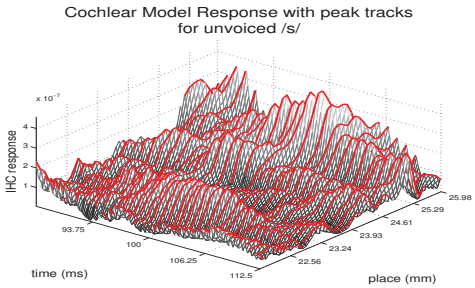

Fig. 4. *Peak tracks from the cochlear response for an unvoiced segment of speech (/s/). The quasi periodic structure that appears in Fig. 3 is not present. Note, that the actual CM response is not plotted for reasons of clarity.*

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 temporally localized distortions.

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, dedicated 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. 2.

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, while the evolution rate is fast for unvoiced sections. It is necessary to track this evolution in both space and time dimensions since the envelope is evolving in both dimensions. Fig. 3 illustrates this evolution for a voiced section of speech by a 3D peak tracking algorithm. It also can be observed that the peak tracks are almost periodic when the rate of evolution is slow as is the case for voiced speech. This parallel structure is lost for unvoiced sections of speech, as is shown in Fig. 4.

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 $0.0684mm/sample$ 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 with Greenwood's map [8] (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 corresponds 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, $p-1$. $t_{forward}$ has a increasing length along place, i.e., faster movement at higher place (lower frequency).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. It is im-

portant 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. 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.

4. Further post-processing involves connecting broken tracks which are possibly 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.

Example of the above steps is illustrated in Fig. 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. Locating Perceptual Formant Regions

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. 3. 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, while 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 Perceptual Formant Regions (PFR), reflecting what is actually been located are 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 Perceptual Formant Regions. This is demonstrated in Fig. 2.

The next step in our feature extraction is to focus on just the Perceptual Formant Regions. 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) are almost constant across the region of each Perceptual Formant Region. This is shown in Fig. 5 which shows that in each of the three Perceptual Formant Regions,all the tracks are paralleled to each other, while the tracks lose this characteristic when they move out of the regions.' Different regions hold with different track distances, which is decided by the place where the formant regions locate.

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 Perceptual Formant Regions, 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 marked as PFR 1, 2 and 3.

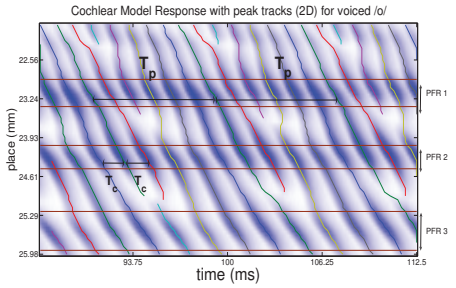

Fig. 5. *Cochlear response with peak tracks for voiced speech /o/ on the time-place plane. The parallel structure between tracks can be observed at the PFRs (between dark horizon lines). Also, the T_c and T_p in Fig. 2 is indicated here.*

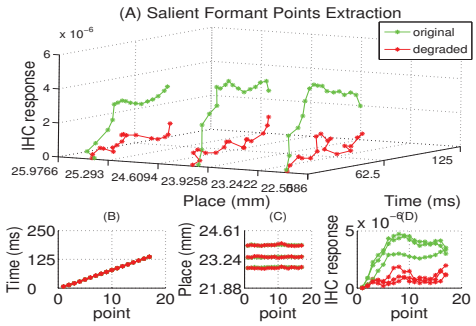

Fig. 6. *Extracted Salient Formant Points. Three perceptual formants are showed. (A) illustrates both original (green) and distorted (red) SFPs as a function of time and place.(B), (C) and (D) shows the time, place and IHC response of the SFPs, respectively.*

2.4. Center of mass for each formant region

A characteristic of the peak tracks in the PFRs is the fact that they are quasi-parallel with a distance of T_c on the time-place plane. The amplitude of peak tracks, however, appear with a period of T_p, instead of T_c, as can be seen from Fig. 2. T_p contains pitch information, which needs to be removed before quality prediction, as the perception of quality is pitch-independent. To extract salient formant information independent of pitch, peak tracks slots restrained by PFR in place are also divided into frames with length of T_p along time. Each set of tracks in a single frame are reduced to a single Salient Formant Point by taking "center of mass" of all the tracks.

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. Finally, Fig. 6.(D) shows the IHC response at each of the extracted points.

3. TEMPORAL ATTRIBUTE OF DAM

SF/SB/SI/SD are four main temporal attributes of DAM, contributing most to Composite Acceptability Estimate (CAE) score [1]. Our

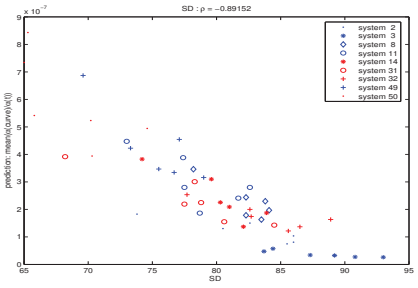

Fig. 7. *J_{rapid} VS Subjective SD Scores*

prediction of these temporal distortions are based on the hypothesis, that such disturbances happen along speakers' formants, where hold most energy and has most significant impact on the listeners perception. Salient Formants Points, as mentioned last section, are used to present perceptual formants, and therefore the predictions of temporal distortions are based on SFPs. A statistical analysis [1] reveals that SF/SB/SI are closer to each other, while SD is slightly far away. Therefore, these four distortions are divided into two categories.

The first category includes SB, SF and SI, which are defined [9, 10] as "Babbling", "Fluttering" and "Interruption" distortion respectively. This long-term category, which means a long-duration, slow evolution distortion over time, has the "slow jitter" computed [11] with formula given below. As an intrusive method, for each degraded SFP_{dis}, corresponding original speech SFP_{ori} is used as the reference of smoothness.

$$J_{slow} = std(|SFP_{dis} - SFP_{ori}|)|_{voiced} \tag{1}$$

The second category contains SD only, defined as "Signal Rasping" and "Crackling" [10], which could be affected by broad range of factors, e.g., center clipping, addictive noise, etc. The difference between SD and the first category is, the former one is a short-time temporal distortion, implying rapid evolution of formants amplitude along time, which leads to the human's feeling of harsh. A formula used to calculate "rapid jitter" for the SD prediction is as below:

$$J_{rapid} = \left.\frac{\partial(|SFP_{dis} - SFP_{ori}|)}{\partial t}\right|_{voiced} \tag{2}$$

4. RESULTS

9 different coding systems were tested, each with three male and three female speakers. Each speech has one prediction score for SB/SF/SI, and another for SD. The correlation coefficients ρ between the subjective DAM scores [9] and corresponding objective predicted scores are calculated for all speech.

For the first category, the predicted score [11] J_{slow} is highly correlated with all three temporal DAM attributes, SB/SF/SI, which are $\rho_{SB,J_{slow}} = -0.91$, $\rho_{SF,J_{slow}} = -0.86$, $\rho_{SI,J_{slow}} = -0.81$.

SD, the only one attribute in the second category, is highly correlated with the prediction of J_{rapid}, which presents the correlation coefficient $\rho_{SD,J_{rapid}}$ of -0.89. Fig 7 reveals the relationship between SD subjective scores and objective predictions J_{rapid}. An improvement can be made by performing monotonic regression [10]. Our test results show that a third order regression can improve the $\rho_{SD,J_{rapid}}$ to 0.93.

5. CONCLUSION

The results above show that the process of extracting the SFP and the subsequent analysis of the IHC deviation is highly correlated with the human perception of temporally localized distortions.Temporal distortion as DAM attributes described, are divided into two categories, SF/SF/SI and SD, based on the evolution changing speed of perceptual formants. An objective "jitter" distortion has been proposed, which has two types, i.e., J_{rapid} for SD attribute prediction, and J_{slow} for SB/SF/SI. The SFPs are closely linked to the formant or the resonant frequencies of the vocal tract and these 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 the cost of computational complexity.

Future work will be focused on the precise prediction of the overall speech quality, i.e., CAE and MOS scores, as a statistical analysis [1] revealed that temporally and frequency localized distortions count up to 70% of variance to the overall speech quality.

6. REFERENCES

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

[2] Joseph L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *JASA*, 2001.

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

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

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

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

[7] D. Sen and J.B. Allen, "Benchmarking a two-dimensional cochlear model against experimental auditory data.," *Midwinter Meeting, Association for Research in Otolaryngology (ARO)*, Feb, 2001.

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

[9] Dynastat, INC., "Diagnostic acceptability measure (DAM): A method for measuring the acceptability of speech over communication systems. Specification DAM-IIC, Dynastat," *Inc., Austin, TX*, 1995.

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

[11] 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.