Locating Discontinuities in Synthetic Speech using a Perceptually Orientated Approach

Joseph Timoney$^{\phi}$, Rudi Villing and Tomas Ward *

$^{\phi}$ Department of Computer Science,
NUI Maynooth
Maynooth, Co. Kildare
E-mail: $^{\phi}$ Jimoney@cs.may.ie

** Department of Electronic Eng.*
NUI Maynooth,
Maynooth, Co. Kildare
E-mail:
**rudi.villing@eeng.may.ie;tomas.ward@eeng.may.ie*

Abstract –**A significant problem with unit selection based speech synthesis is the listener perception of sound discontinuities at which the speech waveforms are joined. This work demonstrates the application of three different perceptually motivated time-frequency representations and associated measures to the identification of such discontinuities.**

Keywords –**Unit selection speech synthesis, Concatenative discontinuities.**

I INTRODUCTION

The current-generation of speech synthesisers employ a waveform-based speech synthesis system that is known as unit selection. Generally, these synthesisers are deemed to produce highly intelligible speech [1]. Furthermore, perceptual naturalness is addressed through the use of pre-recorded speech units. Their principle of operation is to analyse the input text for phonemic and prosodic content, and to then extract the longest possible best -matching sound sequences from a large pre-recorded labelled database of speech that are then joined together to form the output synthetic speech. The selection of best sequences from the database is based on the combination of two costs: target cost (how closely units in the candidate database match the required phonemic targets) and concatenation cost (how well neighbouring units can be joined). Ideally the concatenation cost should correlate highly with human perception of sound continuity. However these synthesisers do not always produce consistent perceptual continuitydue to waveform discontinuities at the concatenation points that will produce an audible non-speech sound. This can be very frustrating for the listener and it detracts significantly from the overall quality of the synthetic speech.

Post-processing of the synthesised waveform can provide a remedy to this problem. Ideally, a signal processing approach would include algorithms that will examine the synthetic waveform and locate points of deviation from perceptual continuity, and then either remove or manipulate the waveform at these points to produce a more natural sounding continuity.

The problem has only recently begun to stimulate relevant research in the area of speech synthesis and so far the most significant result has been in the area of diphone-based concatenation, [2]. Diphone-synthesisers can be viewed as a special case of the unit selection approach, and as such the performance of such a diphone-based synthesiser can be viewed as potentially the worst case performance of a unit-selection based synthesiser. With this work, the perception of discontinuities is attributed to significant formant jumps that can occur at concatenated diphone boundaries. It is suggested that by examining the synthetic signal in the spectral domain the points of waveform discontinuity will be transformed into more clearly observable points of formant discontinuity. It is then hypothesised that where adjacent spectral slices can be shown to differ by a significant amount, then a discontinuity is present in the waveform. To compute the difference between spectral slices, the authors drew on spectral distance measures that are a well-known tool in the area of speech quality assessment. They proposed a measure based on using LPC coefficient descriptions of the spectral envelope whose differences are ascertained using the Kullback-Lieber measure. Following work by [3] which criticised the results of [2] saying the correlation coefficient between their objective measure of discontinuity and the results from subjective tests was too lowa nrew method based on what? Was proopsed. . However, the results of the new method [3] did not offer a significant improvement in the value of the correlation coefficient, obtained(r=0.51).

A difficulty with such a measure as proposed by [2] is that the approach of using an LPC-based spectral distance measure was popular in speech quality assessment in the 1980's [4] but has now been replaced by using measures that are much more pschoacoustically motivated. This changeover has definitely improved speech quality assessment algorithms, for example the PESQ technique has an average correlation coefficient of 0.935 with subjective results [5]. Thus, a similar improvement in the performance of the spectral distance measures applied to concatenated speech should be observable by applying more perceptually orientated spectral distance measures.

II SPEECH SYNTHESISER

The speech synthesis technology used in this work is the CHATR speech synthesis system [6]. CHATR is a unit selection speech synthesiser developed at the speech research laboratories at ATR in Japan. The core architecture is written in C and C++ with the flow of control being specified in LISP, allowing different modules of the program to be selected interactively at runtime. CHATR offers input on many levels. At the most abstract it can accept linguistic descriptions of utterances from which it can generate prosodic phrasing and intonational tune through a rule driven process. Alternatively, the input may explicitly specify prosody and intonation, or, to a further degree, individual phonemes, durations and pitch target values may be defined.

III SPECTRAL REPRESENTATIONS

A common feature of all speech quality measures is that the signals, typically clean and distorted versions, are broken into frames approximately 20msec in length and then converted into the time-frequency domain. At this point, older speech quality measures would take a measure of the spectral envelope of each frame using some transformation of a LPC coefficient description [3]. A difference between the transformed LPC coefficient sets of the two signals would provide the quality metric. More recent speech quality measures have improved by including more psycho-acoustically motivated processing in the signal decomposition and representation process. In general, all methods share three common features:

1. Spectral warping to match the Critical Band resolution of the ear
2. Loudness modelling
3. Masking

Spectral warping implies that in the time-frequency domain the signal energies are warped to match the critical bands of the human ear. These critical bands are frequency selective channels of psychoacoustic processing within the ear, for example, if two tones fall into different critical bands they will be treated by the ear a being almost independent from each other. Loudness Modelling, on the other hand, is a conversion process for each time-frequency slice from being defined on the signal intensity scale measured on the dB SPL scale to one that captures the perceived loudness of each frequency component. The perception of loudness is nonlinear with respect to changes in both frequency and intensity. Finally, masking is a phenomenon by which the threshold of audibility for a sound is raised by the presence of another (masking) sound.

The three spectral distance measures considered in this paper are listed in Table 1.

1. Enhanced Modified Bark Spectral Distortion Measure (EMBSD) [7]
2. Perceptual Evaluation of Speech Quality Measure (PESQ) [5]
3. Perceptual Evaluation of Audio Quality Measure (PEAQ) [8]

Table 1 Spectral Distance Measures investigated

The following sections briefly explain the process of converting the signal into its psycho-acoustic representation using each of the measures.

(i) EMBSD

Once the power spectral density of each Hanning windowed frame is obtained, each power spectrum is partitioned into critical bands of width one bark. The upper frequency limit is assumed to be 3.4 kHz. A spreading function is then applied to this critical band spectrum to take account of the effects of frequency masking across the critical bands using Schroeder's model [7]. After obtaining the spread critical band spectrum the loudness level of each critical band in units of *phon* is obtained using a set of equal-loudness contours taken from the literature and dB intensity values that lie in between the published contours are interpolated to get the correct loudness level [7]. These loudness levels are then converted to sone, the measure of perceived loudness, according to the equations:

$$L(i) = \left(\frac{D(i)}{40}\right)^{2.642}, \quad if \;\; D(i) < 40 \qquad (1)$$
$$L(i) = 2^{0.1(D(i)-40)}, \quad if \;\; D(i) \geq 40$$

.

where $L(i)$ is the perceived loudness of the critical band *i*, and $D(i)$ is the spread critical spectrum in terms of phons in band *i*.

(ii) PESQ

Within the PESQ algorithm the signal is split into overlapping Hanning-windowed time frames and the power spectra is computed [5]. The Bark scale warping is implemented by binning, summing and normalising the power spectrum values into 56 critical bands at a spacing of 0.312 Bark. These are then converted directly to a Sone loudness scale using a formula known as Zwicker's law [9].

$$LX(f)_n = S_l\left(\frac{P_0(f)}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{PPX(f)_n}{P_0(f)}\right)^{\gamma} - 1\right] \quad (2)$$

where S_l is a scaling factor, $P_0(f)$ is the absolute hearing threshold at frequency *f* and $PPX(f)_n$ is the Pitch Power Density at frequency *f* for frame *n*. γ is a 0.23.

Masking is not accounted for by this stage and is only included later in the computation of the actual PESQ value.

(iii) PEAQ

This is the most sophisticated of the psycho-acoustic decompositions [8]. The power spectrum is found of the signal frames after Hanning windowing. The power spectra are weighted by models of the frequency response of the outer and middle ear that is based on a model by Terhardt.

$$A_{dB}(f) = -2.184(f/1000)^{-0.8} + 6.5e^{-0.6(f/1000-3.3)^2} - 0.001(f/1000)^{3.6} \quad (3)$$

The power spectral energies are grouped into Critical bands, spaced at 0.25 Bark An offset is then added to the Critical band energies to compensate for internal noise generated in the ear, again the noise model is adapted from Terhardt.

$$E_{INdB}[f] = 1.456(f/1000)^{-0.8} \quad (4)$$

A triangular (in dB) spreading function is used to implement spreading in the frequency domain.

$$S_{dB}(i,l,E) = \begin{cases} 27(i-l)\Delta z, & i \le l \\ \left[-24 - \dfrac{230}{f_c[l]} + 2\log_{10}(E)\right](i-l)\Delta z, & i \ge l \end{cases} \quad (5)$$

where *i* is the frequency, *l* is the band centre frequency, Δz is the critical bandwidth, and *E* is the excitation pattern.

A time spreading is also introduced that depends on multiple frames. To model forward masking, a frequency dependent filtering or smearing over time made [9].

$$E_f[i,n] = \alpha[i]E_f[i,n-1] + (1-\alpha[i])E_s[i,n] \quad (6)$$

where α controls the time constant of averaging at each frequency, E_f describes initial conditions, and E_s is the excitation pattern.

The model used for loudness calculation is Zwicker's model [9]. However, unlike the PESQ algorithm the values for the excitation threshold in PEAQ they computed using a model description attributable to Kapust [8].

$$LX(f)_n = c\left(\frac{E_t(f)}{s(f)E_0}\right)^{0.23} \cdot \left[\left(1 - s(f) + \frac{s(f)\tilde{E}_{SR}(f)_n}{E_t(f)}\right)^{0.23} - 1\right] \quad (7)$$

where, in terms of dB, the threshold index is given by

$$s_{dB}(f) = -2 - 2.05\tan^{-1}\left(\frac{f}{4000}\right) - 0.75\tan^{-1}\left(\left(\frac{f}{1600}\right)^2\right) \quad (8)$$

the excitation threshold is

$$E_{tdB}(f) = 3.64(f/1000)^{-0.8} \quad (9)$$

and *c* is a constant, set to 1.07664.

III APPLICATION TO SYNTHETIC SPEECH

To test the performance of the synthesiser a number of sources were investigated. Firstly, word tests commonly applied in subjective measurements of speech intelligibility were used. The website [10] provides the word lists of three well-known tests: (1) the Diagnostic Rhyme Test (DRT), (2) the Phonetically Balanced Word List (PB) Test and (3) the Modified Rhyme Test (MRT). These tests combine words, usually having a consonant-vowel-consonant sequence, either in rhyming pairs or embedded in carrier sentences. However, it was found that none of these tests highlighted any significant errors with the speech synthesis system This was attributed to the restricted nature of the sounds and their brief contextual surround. Next, a set of 48 phonetically-balanced sentences provided in [4] that were designed for speech quality tests were synthesised. A number of good examples of discontinuity were obtained. In particular, the synthesised sentence "*the goose laid and odd egg*", with a discontinuity appearing between "odd" and "egg", was selected for testing the usefulness of the psycho-acoustically motivated spectral distance measures.

Careful listening to the sentence followed by a process of making omissions of part of the waveform in the vicinity of these two words was performed to determine the time location of the discontinuity in the waveform. This eventually located the discontinuity to be between 0.75 and 0.8 seconds approximately. Figure 1 shows the region of the waveform identified with the discontinuity in red overlaid over the original waveform, although by visual inspection no striking deformity is observable in the waveform.

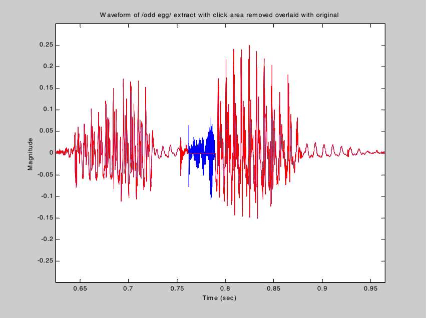

Figure 1 : Waveform with discontinuity removed overlaid in red

To determine the existence of perceptible discontinuities the psychoacoustic decomposition was applied to two signals: (1) the original signal, and (2) a delayed version. The measure would then compute the difference between them. It was assumed that if the time difference between the successive frames under analysis was small then the measured difference should be low as there would be a reasonable degree of spectral continuity over such a short time interval, but only at points of significant spectral difference or discontinuity would a concatenative join be heard by a listener. Following preliminary experiments a window length of 21msec with a frame overlap of 2msec was used. This was judged to be a reasonable compromise, as it was desirable to have a spectral difference measure that would produce low values during periods of normal spectral transitions but that would register the presence of something exceptional like a discontinuity. A long window introduces more averaging in the computation of the spectrum and so adjacent spectral slices will be based on more common information, and thus, their difference should be small. However, if a short component appears that is discontinuous within the typically slowly varying movement of surrounding resonant frequency trajectories, it would show up as being more significantly different by the difference measure.

Figures 2 to 4 show the perceptually motivated time-frequency representations of the waveform segment including the discontinuity as shown in Figure 1 in both red and blue.

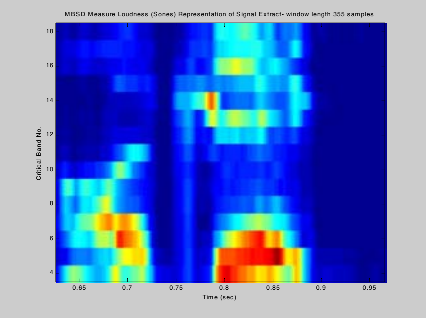

Figure 2 : Perceptual Representation generated for EMBSD Measure

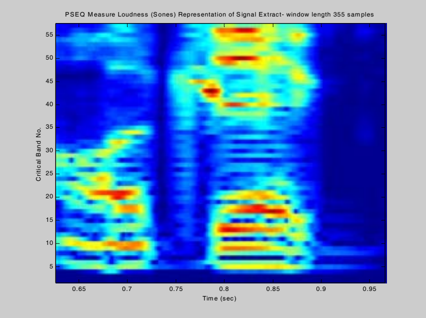

Figure 3 : Perceptual Representation generated for PESQ Measure

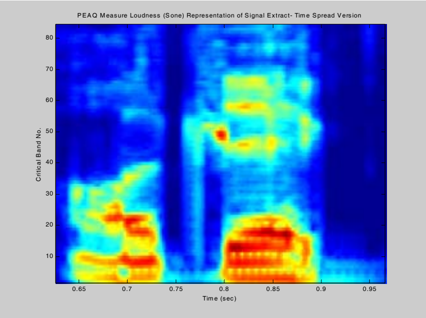

Figure 4 : Perceptual Representation generated for PEAQ Measure

In all the figures unusual activity can be observed at the location of the perceptual artifact (0.77 secs) and in all three cases it appears as a bar that extends from the low frequency region into the high frequency region. In the case of the EMBSD time-frequency representation the peak of this bar is located in the region of 2200Hz. Similarly for the PESQ time-frequency representation, however the spectral bar appears over a slightly greater time interval and is not as clearly defined in terms of its frequency components. The best result in terms of both time and frequency resolution is from the PEAQ time-frequency representation. The greater spectral energy content is contained in the higher frequency region, from about 2315Hz, and extends down towards the low frequency region, around 138Hz. It is also less spread in time that is visible in the PESQ time-frequency representation.

By way of benchmarking, the discontinuity location algorithm of [2] that uses the Kullback-Lieber Spectral Distance Measure was implemented. The results can be seen in Figure 5. From this figure it can be seen that a number of peaks are present, with a large peak at the beginning. This peak does not coincide the time location of the discontinuity, and furthermore, the other that do exist around the appropriate time are of a magnitude that does not indicate any significant spectral differences.

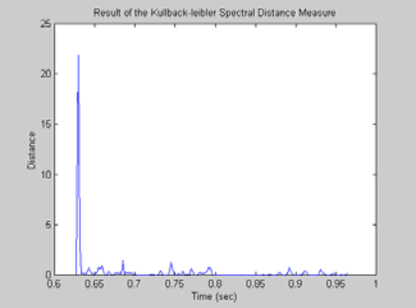

Figure 5 : Result from the Kullback-Lieber Spectral Distance Measure [Klabbers]

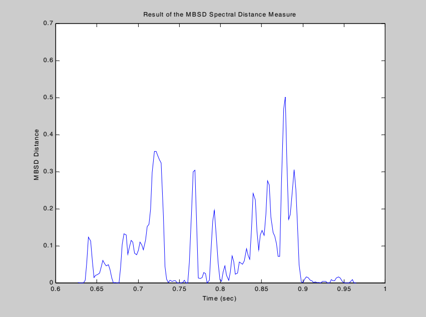

Figure 6 : Result from the EMBSD Spectral Distance Measure

In Figure 6, the result from applying a Euclidean distance measure between consecutive frames of the EMBSD time-frequency representation is given. There is a large peak at approximately 0.77 seconds, the time location of the discontinuity. However, the presence of so many others peaks in the plot make it difficult to create a clear association between this peak at 0.77 sec and the discontinuity. This suggests that this distance measure is not be suitable to the problem in hand.

Figure 7(a) displays the Disturbance Density figure-of-merit of the PESQ algorithm [5]. Again, this measure shows many peaks, with a prominent peak around 0.77 seconds. However, again it is difficult to use this for locating the discontinuity alone as the presence of other significant peaks would confuse any threshold based criterion.. Looking at Figure 7(b), the plot shows the asymmetrical disturbance density figure-of-merit from the PESQ algorithm [5]. This appears to be much less cluttered, two significant peaks only can be seen that are in the vicinity of the discontinuity (0.74 and 0.78 seconds respectively). This is definitely a more desirable answer. It is possible to interpret both peaks as indicating its boundaries, rather than pinpointing the location of the discontinuity itself. This is a more useful result in that it can be used to set thresholds that can be used to make decisions as to where the discontinuity begins and ends in time.

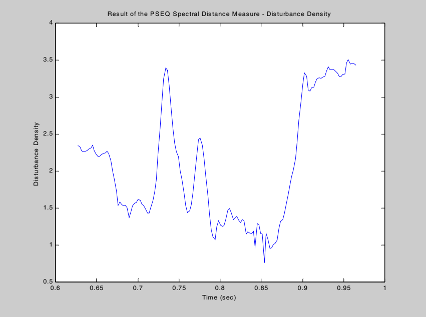

Figure 7(a) : Disturbance Density the PSEQ Spectral Distance Measure

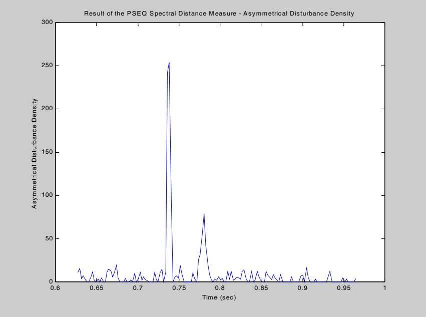

Figure 7(b) : Asymmetrical Disturbance Density the PSEQ Spectral Distance Measure

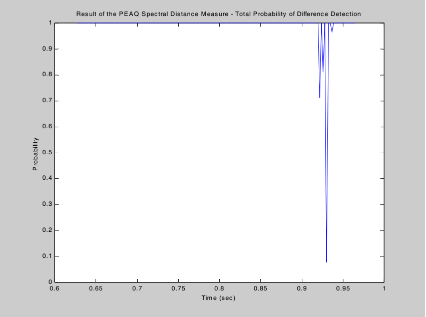

Figure 8 : Results from the Total Probability of Detection of the PEAQ algorithm

Figure 8 shows the results from using the Total Probability of Detection quantity from the PEAQ algorithm [8]. This quantity displays the total probability of the detection of a new component in the signal. However, the probability is 1 for almost the whole segment length, with just a drop in probability value near the waveform end. This means that it is unsuitable for the problem of discontinuity location. Given that the asymmetrical disturbance density showed the best performance with the PESQ, it was decided to apply this measure to the PEAQ psycho-acoustic representation. Figure 9 illustrates this plot. It is not cluttered with spurious peaks, and has significant peaks at the beginning and end of the segment and in the vicinity of the discontinuity: (0.75 and 0.79 seconds). This suggests that asymmetrical disturbance density is a good perceptually-relevant indicator function for concatenative discontinuities. Furthermore, also in comparison to the result for the PESQ measure in Figure 7(b) into account, the PEAQ–based asymmetrical disturbance density in Figure 9 gives a more accurate results in terms of the time location of the discontinuity.

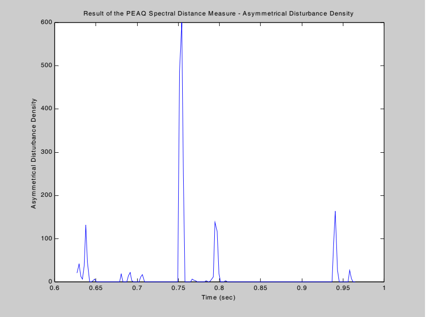

Figure 9 : Results from applying the Asymmetric Disturbance Density to PEAQ time-frequency Representation

IV CONCLUSION

Examining a number of spectral distance measures based on psycho-acoustically motivated time-frequency distributions derived from contemporary speech quality measures, it was found that the PEAQ time-frequency representation combined with the Asymmetrical Disturbance Density measure of the PESQ algorithm gave the best performance in locating the beginning and end points of a discontinuity in concatenated synthetic speech. Future work will be done to determine the appropriate thresholds for this new measure so that it cab be applied automatically to synthetic speech and integrated into the CHATR synthesis system.

V REFERENCES

[1] Keller, E. et al., (eds.), *Improvements in Speech Synthesis*. Wiley and Sons, Chichester, UK, 2001.

[2] Klabbers, E.. Veldhuis, R., ‘Reducing audible spectral discontinuities’, *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, Jan. 2001.

[3] Donovan, R.E., ‘A New Distance Measure for Costing Spectral Discontinuities in Concatenative Speech Synthesisers’, *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Edinburgh, 2001.

[4] Quackenbush, S. R., Barnwell III, T.P., and Clements, M.A., *Objective Measures of Speech Quality,* Prentice-Hall, Englewood Cliffs, NJ, 1988.

[5] Rix, A., Beerends, J., Hollier, M. and Hekstra, A., ‘Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment’, Part II-Psychoacoustic Model, *Jnl. of the AES.*, vol. 50, no. 10. October 2002

[6] Black, A. and Taylor, P., ‘CHATR: a generic speech synthesis system’, *COLING94*, II, Kyoto, Japan, pp 983-986, 1994.

[7] Wonho, Y. et al., ‘Performance of the modified Bark spectral distortion as an objective speech quality measure’, *ICASSP 1998*, vol. 1, 12-15 May 1998, pp. 541 –544.

[8] Thiede, T., Perceptual Audio Quality Assessment using a Non-Linear Filter Bank, Ph.D. thesis, *Technische Universitat Berlin*, Berlin, Germany, 1999.

[9] Zwicker, E., and Fastl, H., *Psychoacoustics, Facts and Models*, Springer, Berlin, 1999.

[10] Word lists, see http://www.meyersound.com/