Speech Quality Evaluation based on AM-FM time-frequency representations

J. Timoney

Department of Computer Science,
NUI Maynooth,
Maynooth,
Co. Kildare,
Ireland
jtimoney@cs.may.ie

J.B. Foley

Department of Electronic and Electrical Engineering,
Trinity College,
Dublin 2,
Ireland

Abstract

This paper deals with the application of information extracted from AM and FM time-frequency representations of speech to the task of determining speech quality. The representations are introduced and then the procedure for data extraction is outlined. The experimental setup for the assessment of objective quality covers distortions typically found in speech communication systems. To determine how well these quality measures perform, regression analysis is used to evaluate how well they estimate the results of subjective testing. Considering each class of distortions individually the objective measures demonstrate good performance, however, this level does not seem to hold as well in the aggregate case. This leads to suggestions as to where possible improvements can be made to the procedure.

1. Introduction

In principle, speech quality should be assessed by subjective methods that rely on listener judgements. However, subjective testing procedures have a number of drawbacks, the most significant being the considerable costs in implementing a suitable program and the variability of the results. To overcome such limitations, a number of objective measures of speech quality have been introduced that attempt to quantify the integrity of the speech signal, most often on the basis of a comparison between the original and a distorted or processed version. The validity of these objective measures is usually made by determining their correspondence with subjective measures via a series of controlled tests. However, developing a good speech quality measure that is applicable over a broad range of distortions is difficult. Over the years, a number of objective quality measures have been suggested, operating either in the time or frequency domain. Generally, results have shown that frequency domain measures correspond better to subjective quality than time domain measures [1].

A well-known class of speech quality measures, known as LPC-based measures, is founded on the linear source-filter model representation of speech. The inherent simplicity have ensured their popularity but they are known to produce inaccurate and defective results due to limitations in the model [2]. One outstanding deficiency is that these methods fail to take account of both perceptually important dynamic changes occurring in the speech [3] and the finer spectral details. By way of an alternative, this paper proposes a novel approach that examines speech quality from an Amplitude Modulation-Frequency Modulation (AM-FM) perspective. In recent years, there has been much interest in the amplitude and frequency modulation structure of speech as it attempts to overcome previous deficiencies in speech modelling by indirectly describing the non-linear and time-varying phenomena that occur during speech production [4]. It is also motivated by better understanding of the signal processing function performed by the inner ear, particularly the cochlea. The cochlea is known to decompose acoustic stimuli into frequency components along the length of the basilar membrane. This phenomenon is called Tonotopic decomposition. Further it is also known that the nerve fibres emanating from a high-frequency location in the cochlea "phase-lock" to the envelope of the stimulus around that frequency, i.e. convey information about the envelope modulations in the signal. Thus, to a first-order approximation, it is often argued that the tonotopic location/place along the length of the basiliar membrane conveys the FM or frequency information about the signal, and the rate of nerve fibre activity around that location conveys the AM or envelope information [5].

In regards to the design of an objective measure of speech quality that takes account of these phenomena, this investigation uses information from separate time-frequency representations that display the most significant features of the AM-FM speech structure. In particular, the Modulation Spectrogram [6] and the Instantaneous Frequency Distribution [7] are employed to detail the respective AM and FM components of the speech. By considering both aspects, inclusion of elements of the finer spectral details with the dynamic features of the spectral envelope is ensured. Thus, the objective quality measure will encompass a perceptually relevant set of speech features and should provide a better match to subjective results.

2. AM-FM Time-Frequency Representations

This section outlines the signal processing procedures required to form the Modulation Spectrogram and Instantaneous Frequency Distribution of a speech signal. An example is also given of each to illustrate the features of the speech signal that they emphasise.

2.1 Modulation Spectrogram

This representation was proposed in response to the significant evidence showing that much of the phonetic information of the speech signal is encoded by slow changes in gross spectral structure that characterise the low-frequency portion of the amplitude modulation spectrum. One excellent example of this is the Channel Vocoder [8], another is from Houtgast and Steeneken [9] who, in the development of the STI speech quality test, established that there was a significant connection between speech quality and the attenuation of the low-frequency modulation components present in the signal. Thus, the modulation spectrogram was developed as a time-frequency representation to display the short-term low-frequency modulations present in the amplitude envelope of the speech. These low-frequency modulations are exposed across critical-band channels to further enhance the perceptual relevance of the representation [6].

To generate the Modulation Spectrogram, the incoming speech is separated into critical-band-wide channels using a Mel-spaced FIR filter bank; in this work 19 channels are used. To ensure that there is minimal overlap between the channels, the critical-band filters are designed to have a twenty-fourth octave bandwidth. In each channel, the signal envelope is derived using a Hilbert transformer and is normalised[1] by its power, filtered with a 50 Hz lowpass filter and then decimated by a factor of 100. In [6], the low-frequency modulations present in these decimated envelope signals were found by computing a 128-point FFT over a 250 ms Hamming window which is updated every 12.5 ms in order to capture their dynamic properties. However, an alternative to the FFT for the computation of the desired low frequency modulation spectrum of the decimated envelope signals is to use the Chirp z-transform [10]. This algorithm determines samples of the z-transform of the signal along an equally spaced spiral contour defined between two desired frequency points. It has the advantage of being able to calculate the samples with an arbitrary starting point and frequency range, and it allows calculation of an arbitrary number of samples along this contour thus reducing the error in frequency representation. One potential drawback is that, depending on the number of sample points, it may be slower than the FFT. Finally, the power of the modulation components from 0-8 Hz in each channel is then plotted in spectrographic format, and for visual enhancement both thresholding and smoothing can be applied.

Figure 1 shows a Modulation Spectrogram of the utterance "Do not present the prize". The darker areas in the figure show greater intensities of low frequency modulations, appearing most noticeably at the output of the channels in the proximity of 1000Hz. Also, the dynamics of the transitional events in the utterance are highlighted.

[1] By this normalisation, the *intensity* envelope is actually obtained. This was found to give a superior peak resolution and visual representation than if the normalisation was performed by subtracting the DC component from the decimated envelope signals as in [6].

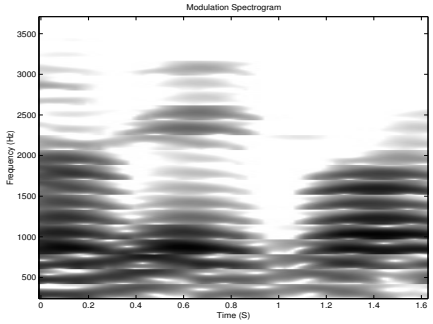

Figure 1: Modulation Spectrogram of a Speech Utterance

2.2 IF Distribution

In keeping with the AM-FM model of speech, a complementary representation to that of the Modulation Spectrogram is one that displays the underlying FM structure of the signal. The first significant investigation into a representation of the FM or instantaneous frequency structure of the speech signal was carried out in [7]. Experiments using the Phase Vocoder have demonstrated the perceptual relevance of the instantaneous frequency structure of a speech signal [10]. The representation suggested by [7] is achieved through an FFT based signal processing scheme where the running short-time Fourier transform (RSTFT) is interpreted in terms of simultaneous outputs of a bank of band-pass filters with successively offset centre frequencies, all having the given signal waveform as input. This is defined as

$$X_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} x(n+m)w(m)e^{-j\omega m} \qquad (1)$$

where $w(m)$ is a real "window" sequence which determines the portion of the input signal which receives emphasis at time *n*. The local instantaneous frequency at any ω and *n* can be obtained from

$$v(\omega,n) = \mathrm{Im}\left\{\frac{1}{X_n\left(e^{j\omega}\right)}\frac{\partial}{\partial n}X_n\left(e^{j\omega}\right)\right\} \qquad (2)$$

A convenient method for computing $v(\omega,n)$, without explicit differentiation of the actual time functions, can be derived by manipulation of (1) and (2) to give the Instantaneous Frequency Distribution (IFD) across time

$$v(\omega,n) = \omega + \mathrm{Im}\left\{\frac{\sum_{m=-\infty}^{\infty} x(n+m)\,w'(m)e^{-j\omega m}}{\sum_{m=-\infty}^{\infty} x(n+m)\,w(m)e^{-j\omega m}}\right\} \qquad (3)$$

where

$$w'(n) = -\frac{\partial}{\partial n} w(n)$$

When calculating the IFD for speech signals of greater duration than 1 second, it is desirable to introduce a running short-time averaging step to the data contained in the IFD to aid clarity of presentation. Assuming that the phonetic content of the speech signal is adequately represented by a vector frame interval of 10 to 20 msec, averaging over this time period should not significantly affect the nature of the data. One approach [11] is to weight the IFD by the corresponding squared amplitude before averaging, which can be conveniently expressed as

$$\bar{v}(\omega, m) = \mathrm{Im}\left\{ \frac{\sum_{l=n_0}^{n_0+N} X_l'\left(e^{j\omega}\right) X_l^*\left(e^{j\omega}\right)}{\sum_{l=n_0}^{n_0+N} X_l\left(e^{j\omega}\right) X_l^*\left(e^{j\omega}\right)} \right\} \qquad (4)$$

where $X_n'\left(e^{j\omega}\right)$ is the time derivative of $X_n\left(e^{j\omega}\right)$ as given by (1), $X_n^*\left(e^{j\omega}\right)$ is the complex conjugate, N is the length of the averaging window and m is the frame index.

Figure 2 shows a plot of the average weighted IFD calculated with (4) applied to the sentence "You are the biggest man" where averaging was performed over a 10 msec frame interval with a frame update of 5 msec.

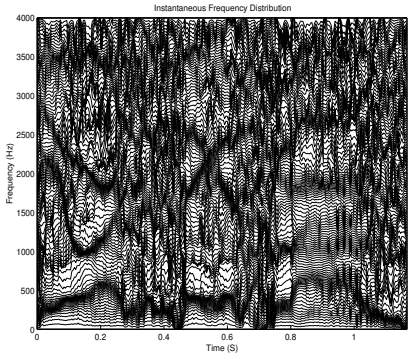

Figure 2: Average Weighted IFD of an Utterance

The formant locations are readily seen in the figure as dense areas of concentration in a similar way to the high Fourier amplitudes that outline the formant tracks in the speech spectrogram. However, in the IFD, the resolution of the formants is clearer and the tracks of the higher frequency formants are more easily distinguished. Areas of low density can also be seen which possibly represent spectral zeros [7].

3. Time-Frequency Distance Measures

To quantify speech signal distortions using the AM-FM based time-frequency representations described above some form of processing to parameterise the information is desirable. A possible solution is to convert the information at each instant in the time-frequency plane into a set of Cepstral-like coefficients which will thus allow a degree of data reduction that will diminish the complexity of the processing procedure [11]. Each time-frequency slice is concatenating a mirrored replica of itself so as to define it up to the speech sampling frequency. As in conventional homomorphic processing, the natural logarithm of this concatenated section is taken, followed by an inverse FFT to transform it to the Cepstral domain. A suitable number of Cepstral coefficients can then be extracted depending on the required accuracy or smoothness. For the IFD, however, it is necessary to convert the densities that appear in the plot into a form that indicates the level of concentrations at each frequency bin. This can be done by forming a histogram of the time-frequency slice, where the maxima in each histogram represent spectral areas of dense component concentration. Simple thresholding can also be applied to these histograms to remove peaks indicating areas of small concentrations. Experimental evaluation found that 40 Cepstral coefficients were sufficient in both cases to describe the information in each time-frequency slice. The distance measures for each time frequency representation, D_{MS} and D_{IFD} can then be taken as the average mean square difference of the Cepstral coefficients for the original and distorted speech waveforms over all the speech frames. This can be written as

$$D = \frac{1}{M} \sum_{m=0}^{M-1} \bar{e}^2(m) \qquad (5)$$

where $\bar{e}^2(m)$ is the mean square frame difference and D, the distance value, is the average of $\bar{e}^2$ over the total number of frames M.

3.1 Experiments

To test the AM-FM speech quality measure, a speech database was created using the list of 48 phonetically balanced sentences provided in [2]. The sentences were recorded in four separate groups using two male and two female speakers. The types of distortion applied to the speech in the tests were based on frequently occurring distortions in communication channels [3]. These were applied in various degrees, resulting in a total of 18 conditions, as specified in Table 1.

Distortion	Degree
Masking noise	SNR= 35, 25, 15, 10, 5 dB
Peak clipping	7,30,50,70, 90 % (cut-part/whole)
Band-pass filtering	0.8-1.3,1.3-1.9,1.9-2.6,1.4-3.2 kHz
Single echoes	1.25ms (reflection coefficient 0.5 and 0.6), 6.25 and 12.5 ms (reflection coefficient 0.5)

Table 1: Distortions Applied to the Speech

Subjective assessment of the distorted speech signals was made using the Mean Opinion Score (MOS). To assess how well the distance values could predict the subjective results, a Multiple Linear Regression procedure [2] was used. To validate the procedure the data was split in two, the first half being used to estimate the regression weights and the second to compute the figure-of-merit, here the correlation coefficient. Table 2 shows the results of this procedure.

It is clear that when the distance measures are examined for each individual distortion, there is a very high correlation between the subjective and objective results. However, when the correlation is calculated over all the distortions, it is not so high. Still, a correlation coefficient of 0.75 sets it above those found for the class of LPC-based measures examined in [2]. It seems that the quality measure works least well in the case of echo distortion. This could be attributed to the fact that although an echo may cause significant spectral deformities, the perceptual quality of the speech may not suffer proportionally.

Distortion	Correlation Coefficient
Masking noise	0.9974
Peak clipping	0.9874
Band-pass filtering	0.9785
Single echoes	0.9479
All distortions	0.7509

Table 2: Correlation Coefficients between Objective and subjective Results for Various Distortions

4. Conclusion and Future Work

In conclusion, this paper has proposed a methodology has for the construction of a speech quality measure based on AM-FM time-frequency representations of speech. This measure was shown to correlate very well with subjective results for each individual class of distortion, however, this performance diminished when considering its correlation with the aggregate distortion set.

A number of suggestions can then be made that could improve the performance of the AM-FM distance measure. The first would be to enhance the Instantaneous Frequency distribution by making it more perceptually relevant. One approach to achieve this would be to warp the speech segments in time before taking the FFT as proposed in [12]. Another would be to take account of the importance of human judgement in speech perception by employing the Measuring Normalising blocks procedure described in [13]. Finally, better correlation with the subjective results could be obtained by using an alternative methodology to combine the distance values from both distributions, such a Neural Network or Multivariate Polynomial Regression [14].

5.0 References

[1] N. Kitawaki et. al, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE Jnl. Selected Areas Comms.*, , Vol.6, No. 2, Feb. 1988, pp. 242-247.

[2] S.R. Quackenbush, T.P. Barnwell and M.A. Clements, *Objective measures of speech quality*, Prentice Hall , Englewood Cliffs, NJ, 1988.

[3] S. Wu and L. Pols, "A distance measure for objective quality evaluation of speech communication channels using also dynamic spectral features," *IFA Proc.*, Vol. 20, 1995, pp. 27-43.

[4] A. Potamianos, '*Speech processing applications using an AM-FM modulation model*,' PhD. Thesis, Harvard University, Massachusettes, MA, Aug.1995.

[5] A. Rao and R. Kumaresan, "Effect of temporal envelope smearing on speech reception," *Journl. Acoust. Soc. Amer.*, vol. 105, no. 3, March1999, pp. 1912-1924.

[6] S. Greenberg and B. Kingsbury," The modulation spectrogram: In pursuit of an invariant representation of speech", *ICASSP 1997*, Munich, Germany, pp. 1647-1650.

[7] D. Friedman, "Instantaneous frequency distribution vs. time: an interpretation of the phase structure of speech," *ICASSP 1985*, pp. 1121-1124.

[8] Homer Dudley, "Remaking Speech," *Journl. Acoust. Soc. Amer.*, vol. 11, no. 2, Oct.1939, pp. 169-177.

[9] T. Houtgast and H. Steeneken, "A physical method for measuring speech-transmission quality", *Journl. Acoust. Soc. Amer.*, vol. 67, no.1, Jan. 1980, pp. 318-326.

[10] L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[11] D. Friedman, "Formulation of a vector distance measure for the instantaneous-frequency distribution (IFD) of speech," *ICASSP 1987*, pp.1748-1751.

[12] J. Garas and P. Sommen, "Warped linear time invariant systems and their application in audio signal processing," *ICASSP 1999*, pp.1209-1213.

[13] S. Voran, "Objective estimation of perceived speech quality using measuring normalising blocks," NTIA Report 98-347, US Dept. of Commerce, April 1998.

[14] H. Wang and D. Vaccari, "Multivariate polynomial regression for identification of chaotic time series," http://attila.stevens-tech.edu/~dvaccari/nature.html.