Autoregressive Time-Frequency Interpolation in the Context of Missing Data Theory for Impulsive Noise Compensation

I. Potamitis, N. Fakotakis

Wire Communications Laboratory, Electrical and Computer Engineering Dept., University of Patras, 261 10 Rion, Patras, Greece, Tel:+30 61 991722, Fax:+30 61 991855
potamitis@wcl.ee.upatras.gr

Abstract
The present paper reports on a novel technique for the identification and replacement of spectral coefficients degraded by impulsive noise. The problem is viewed from the perspective of Missing Feature Theory (MFT). The analysis is carried out in the linear spectrum prior to, or after applying the mel-scale filter-bank depending on whether one aims at improving the quality of perception of speech recordings or at Automatic Speech Recognition (ASR). Each filter-bank output is considered to be a time series drawn from an Auto-Regressive process (AR). A validation corpus of undistorted recordings is used to derive a-priori bounds on the expected prediction error of each AR model. In operational conditions, the prediction procedure is monitored and the violation of the statistical bounds indicates band corruption and entails the substitution of the degraded spectral coefficients by the prediction of the corresponding AR model. ASR experiments and informal listening tests demonstrate large improvement in terms of word recognition performance and Itakura-Saito divergence at very low Signal to Impulsive Noise Ratios (SINRs). Data, and implementation code can be found at: ftp://wcl.ee.upatras.gr

1. Introduction

ASR systems are composed of a feature pre-processing stage, which aims at extracting the linguistic message while suppressing non-linguistic sources of variability, and a classification stage (including language modelling), that identifies the feature vectors with linguistic classes. The extraction level of current ASR systems converts the input speech signal into a series of low-dimensional vectors, each vector summarizing the temporal and spectral behaviour of a short segment of the acoustical speech input. The ultimate goal is to estimate the sufficient statistics to discriminate among different phonetic units while minimizing the computational demands of the classifier. The classification stage that is based on the probability density function of the acoustic vectors is seriously confused in the case of a mismatch between training and operational conditions (we assume an HMM classifier). MFT makes the assumption that the part of the spectrum that is degraded by ambient noise to the point that it diverges from the spectral characteristics of the speech waveforms used to train the classifier is useless for the recognition purpose. Therefore, the time-frequency representation is decomposed to reliable regions (regions that match the training representations) as well as unreliable regions. From this point, there are two different ways that the stochastic framework of continuous density Hidden Markov Models (HMMs) can be adapted to handle time frequency regions corrupted by noise.

The first carries out the classification process based on the degraded observations ignoring the existence of the unreliable data, namely the marginalization process. The second category, namely the data imputation method, infers the unreliable data that have been conditioned on the degraded ones. There are many potential advantages to the MFT approach with regard to recognition of speech impaired by noise [1]-[6]. Regardless of the view to encounter the unreliable features, MFT is based on the robust identification of the unreliable spectral coefficients over time. To our knowledge, three criteria are implemented in the context of MFT theory to serve the identification purpose. The first one employs spectral subtraction and treats the regions that attain negative values after subtractions as unreliable. The second is based on the local estimation of SNR and the characterization of parts of the time-frequency representation that possess SNR under a certain threshold as unreliable [1][2]. A third approach is presented in [5] where a classifier-based estimation of spectrographic masks is treated as a problem of Bayesian classification.

Our framework introduces a novel solution to the problem of identification of degraded spectro-temporal regions, which is based on the prediction error of the AR process that models each spectral band. We look the outputs of each filter bank as a separate informational stream and we fit a nine order AR model to predict the next spectral coefficient in each band. Subsequently, we apply the prediction procedure to each band of a validation set that consists of recordings clean from any kind of impulsive noise, and measure the Mean Squared Error (MSE) between the predicted spectral coefficient and the actual one. The error measurements are used to derive the expected error bounds of the prediction error for each band in order to form confidence intervals on a new observation from the process. The objective is to use statistical process control to monitor the prediction error in case it changes in an undesirable systematic direction in operational use. In operational use, if the mean square error between the predicted spectral coefficient and the actual one falls within the control limits, the spectral coefficient is classified as clean and remains intact. In case of an impulsive distortion, the AR model produces an erroneous prediction outside the statistical control limits calculated from the validation set indicating that the corresponding spectral coefficient of the next frame is far from being expected. In the latter case, the following spectral coefficient is classified as an outlier and is subsequently replaced by the prediction. We present its application to a highly non-stationary noise type, which consists of short-time bursts of random amplitude, spectral content, onset time and frequency of occurrence. It can deal with heavily distorted speech that may extend to several hundred samples. The sharp changes in the prediction error that are indicative of band
corruption are attributed to the fact that impulsive noise intrusion is short time and the AR models base their prediction on the previous undistorted spectral coefficients.

Our approach possesses some distinct advantages such as:

a) The noise suppression scheme avoids a global act on the original waveform, that is, the restoration procedure is activated whenever corruption is detected and is applied locally, therefore, it does not inflict any distortions on an already clean part of the spectrum.

b) Since we model each band with a different AR model, we can deal effectively with band limited impulsive corruption.

c) Our method belongs to the imputation category, which has the advantage over marginalization that it does not constrain the recognizer to the log-spectrum domain. After restoration we can proceed in applying decorrelation techniques (e.g. Discrete Cosine Transform (DCT), Karhunen Loeve Transform), and expand the static vector by adding deltas and double deltas) that are well known to improve recognition.

2. Missing Data Estimation in AR Series

We introduce the problem of degradation of the time domain speech signal \( \{ s_t \} \) where \( t \) denotes the sample index, due to the presence of additive impulsive noise. Impulsive noise appears in the form of a sequence of short pulses each having length \( M_{i} \) samples, where \( \{ i = 1, \ldots, P \} \) denotes the index number of each distinct pulse. Let \( \{ t \} \) be the random onset time of the pulse \( \{ i \} \). The signal \( s_{t} = [ s_{t}, \ldots, s_{tM_{i}-1}], \ldots, [ s_{t}, \ldots, s_{tM_{i}-1}], \ldots, [ s_{t}, \ldots, s_{tM_{i}-1}] \) is assumed unobserved at sampling points \( \{ s_{t}, \ldots, s_{tM_{i}-1} \} \), \( \{ i = 1, \ldots, P \} \) where intrusive pulses appear. There are a number of interpolating techniques for the replacement of a sequence of samples in time domain. Namely, Least Squared Autoregressive Interpolation and its modifications [7], template-based methods [8] and Gibbs sampling framework [9] are efficient techniques for compensating the effect of a small number of samples. The disadvantage of these methods becomes apparent in degradation of several hundred samples as they appear in real adverse channel environments, e.g.switching noise, slumming doors, etc. Markov Chain Monte Carlo techniques [10] are more powerful for the interpolation of large sequence of samples, but, at the time being, they are computationally quite intensive to be used in ASR.

Our technique aims primarily to ASR, therefore, we retain the principles of MFT and we place them in a framework with low computational requirements. Cepstral coefficients have proven more effective than spectral or log spectral representations in the context of ASR. However, cepstral feature vectors make use of information from all spectrum bands; therefore, any distortion induced to any part of the spectrum is spread to all features forming the vector. Our technique identifies the corrupted spectro-temporal regions of the speech signal, discards its samples and substitutes the missing gap based on the weighted average of a number of uncorrupted samples prior to the ones to be restored. We implement the restoration procedure prior to the DCT of the Mel Frequency Cepstral Coefficients (MFCC) feature extraction stage, and after the application of the mel-scale filter-bank. We also present results for speech quality improvement. In the later case the algorithm is applied to the linear spectral bands of the Short Time Fourier Transformed time domain signal. The appeal of the time-frequency domain instead of the time domain lies in three reasons:

a) The problem of interpolating \( N \) samples in time-domain can be converted to a problem of substituting one sample across time for each FFT band [8].

b) There is extreme redundancy of information in the time-spectrum domain, a fact that we can exploit in order to build robust interpolators that restore the unreliable acoustic features based on the reliable parts of the spectrum. Moreover, part of the success of any interpolation procedure is due to the correlation of successive samples that can be utilized in order to infer the missing samples. Consecutive spectral coefficients belonging to a spectral band are strongly correlated.

c) As regards noise compensation in the context of ASR, most of the feature extraction front-ends of contemporary ASR machines employ in their construction an FFT stage. Avoiding employing transformations other than the already extracted for the recognition purpose, the function of the outlier identification and restoration model comes at small computational overhead.

2.1 Impulsive Noise Identification

The correlation structure of each band is independently modelled by an AR process, thus, we have implicitly assumed independence among frequency components. This is actually an approximation since adjacent frequency bands are correlated. Let \( \{ n \} \) denote the band either in the linear spectra or mel-filterbank domain. The generative process describing the inference of a spectral coefficient at time sampling instant \( \{ t \} \) in band \( \{ n \} \), is stated in Eq. 1. Context information is incorporated through the vector \( x_{a} = [ x_{a1}, \ldots, x_{ap} ]^{T} \) and is weighted by the coefficients \( a_{a} = [ a_{a1}, \ldots, a_{ap} ] \).

\[
x_{a}(t) = \sum_{k=1}^{p} a_{ak} x_{n}(t-1) + e_{a}(t)
\]

where \( e_{a} \) is assumed to be white, Gaussian excitation sequence with variance \( \sigma_{e}^{2} \).

Let \( x_{a}(t) \) denote the original spectral/mel-filterbank coefficient corresponding to each frame of a validation set that consists of an ensemble of 500 phonetically balanced recordings. We define the divergence of the prediction \( x_{a} \) from the true \( x_{a}^{\wedge} (t) \) as:

\[
E_{a}(t) = x_{a}(t) - x_{a}^{\wedge}(t)
\]

In order to assess the extent of the expected divergence for each band we need a measure of the distortion. We have used mean square error between the true and predicted value over each band of the ensemble defined as:

\[
D_{a} = \frac{1}{files} \sum_{f=1}^{files} \sum_{frames} \sum_{n=0}^{\text{frames}-1} (x_{a}(t+m) - x_{a}^{\wedge}(t+m))^{2}
\]

and the expected divergence of the prediction error:

\[
D_{s} = \frac{1}{files} \sum_{f=1}^{files} \sum_{frames} \sum_{n=0}^{\text{frames}-1} (x_{a}(t+m) - D_{a})^{2}
\]

The derivation of the error bounds is time-consuming especially when we attempt at improving the quality of speech perception. On the other hand they have to be calculated only once for all subsequent restorations. A short number of experiments did not prove any strong connection between the number of recordings used to derive the error bounds and inference precision.
2.2 Impulsive Noise Compensation

In operational use, if the mean square error between the predicted value and the actual one falls within $[0, D + \alpha D_{\text{out}}]$, where $\alpha$ is an empirically tuned threshold the feature is classified as clean and remains intact. In the case of a degraded spectral coefficient in band $[n]$, practical tests, indicated that the prediction error of an impulsively degraded spectral coefficient is at least one order of magnitude higher than the expected error, therefore is classified as impaired and is subsequently replaced by the predicted spectral coefficient. The identification of the outlier is attributed to the fact that the AR model bases its prediction on spectral coefficients clean of any kind of impulsive corruption as indicated by the annotated transcription of the annotated speech-database that was used. One can utilise a predictive or a probabilistic model of the signal make an optimal estimate of the missing segment using samples on both side of the gap. We did not apply this technique on the grounds of short system delay and on-line performance (the processed frame lags one frame behind the last input).

![Fig. 1. Schematic diagram of impulsive noise identification and compensation front-end.](image)

3. Simulation and results

Due to the extreme departure of impulsive noise from the stationarity assumption, the classical Signal to Noise Ratio (SNR) concept must be adapted. We follow the notation of [8] regarding SNR measurements in the presence of impulsive noise. The entire noise sequence depends on two parameters: a) The average power of each pulse, b) the number of distinct impulses or equivalently the rate of occurrence of incoming impulses. Let $\{P_{\text{im}}\}$ denote the average power of each impulse and $\{P_{\text{sig}}\}$ the signal power. An average signal to impulsive noise ratio is defined in Eq. 5 where $\langle \alpha \rangle$ is the fraction of signal samples contaminated by impulsive noise as:

$$\text{SNR (dB)} = 10 \log_{10} \frac{P_{\text{sig}}}{\langle \alpha \rangle P_{\text{im}}}$$

(5)

In order to carry out an efficient evaluation process, we must design an experimental procedure that captures the effect of the random occurrence, rate, and spectral content of impulsive noise. We extracted a pulse of 500 samples duration from Machine-gun noise down-sampled at 8KHz from the NOISEX-92 database. For each recording of the 100 clean record set, we construct a noisy signal having a varying number of pulses, at random onset time, so that the percentage of distortion samples is matched according to Table 1. Subsequently the noisy signal is artificially added to each recording in five different SNRs so that the corrupted waveform ranges from -10 to 20 SNRdB. To be more specific; let $y_1$ be a clean recording and $y_2$ the noise waveform. The resulting noisy recording $y$ is $y = y_1 + h \ast y_2$, such that $10 \log_{10}(E[y_1]/E[h \ast y_2]) = \text{SNR}_{\text{dB}}$ where $E[.]$ denotes energy. The objective criteria are based on the mean value over all recordings.

3.1. Spectrum Restoration

One should note in Fig. 2 that the fine structure of the formants is preserved and there is almost no sign of visually perceptible distortion. In order to objectively assess the extent of the divergence between the clean and enhanced signal we employed the Itakura-Saito (IS) distortion measure (Fig. 3). IS, is very sensitive to speech spectrum variations but not to phase distortion and is broadly accepted to be well associated with speech quality assessment. It is based on the spectral distance between AR coefficient sets of the clean and enhanced speech waveforms over synchronous frames of 15ms duration. IS distortion is zero in complete match of the enhanced signal with the original waveform. As can be observed the remaining distortion is very small even at low SINRs and high occurrence rates and in good agreement with theoretical expectations.

![Fig. 2. Impulsive noise corruption at 30% of the samples of the original waveform at 0 dB. Row a,b) Corrupted signal in time and in Time-Frequency domain, c,d) Enhanced signal.](image)

![Fig. 3. Itakura-Saito distortion measurements](image)
3.2. Mel-Bank Restoration

For the evaluation of our system we used a speech recognition module built with HTK Hidden Markov Models toolkit [11]. The basic recognition units are tied state context dependent triphones of five states each. In order to train the reference recogniser we used the SpeechDat-ILGR database of utterances and their associated transcriptions. The baseline performance is 96.07%. The testing set is comprised of 100 files, part of the identity card corpus of the SpeechDat database (one speaker for each recording). Each input speech signal waveform is sampled at 8kHz, pre-emphasized by the filter \( H(z)=1-0.97z^{-1} \) and subsequently, windowed into frames of 20ms duration at a frame rate of 10 ms using a Hamming window. Each frame processed so far is Fourier transformed and then passed through a set of 20 Mel-spaced triangular band-pass filter-bank channels. The enhancement process is applied and subsequently, DCT is applied to 20 log-filter-bank channels which reduces the feature vector into 12 dimensional MFCC features plus a log-energy value. Cepstral mean normalization was applied. Deltas and double-Deltas were concatenated to form the final 39-dimensional observational vector. Table 1 depicts the average word recognition results, which demonstrate a clear and considerable gain when the enhancement procedure is included in the MFCC front-end. Recognition performance degrades mainly because of the high number of occurrences and not as a result of high-energy pulses. The system fails only in cases of large clusters of adjacent, high energy disturbances where the prediction is possible use with stationary and quasi-stationary coloured noise types. We are currently constructing a variation of our framework is based on the identification of spectral vectors which are not drawn from the underlying spectrum of the speech signal and which are treated as outliers. We suggest that our method can be used in conjunction with noise compensation techniques that deal with stationary types of noises. As regards its utilization in ASR, it does not inflict any significant computational overload; a property that makes it suitable for handling real-world impulsive corruption. Future developpment of our framework will attempt to deal with its no specific characteristics of the machine-gun noise since our method can be used in conjunction with noise compensation techniques that deal with stationary types of noises. As regards its utilization in ASR, it does not inflict any significant computational overload; a property that makes it suitable for handling real-world impulsive corruption. Future development of our framework will attempt to deal with its possible use with stationary and quasi-stationary coloured noise types. We are currently constructing a variation of our algorithm, which will lift the assumption of band independence by exploiting Multivariate Autoregressive Models (MARs).

4. Conclusion

Our technique has been investigated with a view to introduce a novel measure capable of identifying unreliable spectral regions and to place this measure in the framework of MFT. This technique finds application in the replacement of medium to large gap of missing samples corrupted irrevocably by impulsive noise in the context of ASR and of speech perception improvement. The objective is to interpose an identification stage responsible for the detection of corrupted spectral coefficients based on AR modelling of the spectral pattern of each spectral band and a restoration stage that acts locally on the degraded spectrum. We have not made use of any specific characteristics of the machine-gun noise since our framework is based on the identification of spectral vectors which are not drawn from the underlying spectrum of the speech signal and which are treated as outliers. We suggest that our method can be used in conjunction with noise compensation techniques that deal with stationary types of noises. As regards its utilization in ASR, it does not inflict any significant computational overload; a property that makes it suitable for handling real-world impulsive corruption. Future development of our framework will attempt to deal with its possible use with stationary and quasi-stationary coloured noise types. We are currently constructing a variation of our algorithm, which will lift the assumption of band independence by exploiting Multivariate Autoregressive Models (MARs).

5. References


<table>
<thead>
<tr>
<th>SINR</th>
<th>-10</th>
<th>-5</th>
<th>0</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-En</td>
<td>92,14</td>
<td>93,57</td>
<td>95,71</td>
<td>96,07</td>
<td>96,07</td>
</tr>
<tr>
<td>Enh</td>
<td>95,38</td>
<td>95,82</td>
<td>96,00</td>
<td>96,07</td>
<td>96,07</td>
</tr>
<tr>
<td>No-En</td>
<td>86,43</td>
<td>92,14</td>
<td>93,21</td>
<td>95,29</td>
<td>94,64</td>
</tr>
<tr>
<td>Enh</td>
<td>93,82</td>
<td>94,94</td>
<td>94,94</td>
<td>96,07</td>
<td>96,07</td>
</tr>
<tr>
<td>No-En</td>
<td>84,29</td>
<td>89,64</td>
<td>91,09</td>
<td>94,26</td>
<td>96,07</td>
</tr>
<tr>
<td>Enh</td>
<td>90,91</td>
<td>94,15</td>
<td>92,13</td>
<td>96,07</td>
<td>96,07</td>
</tr>
<tr>
<td>No-En</td>
<td>80,36</td>
<td>85,86</td>
<td>91,01</td>
<td>93,10</td>
<td>96,07</td>
</tr>
<tr>
<td>Enh</td>
<td>89,01</td>
<td>89,01</td>
<td>92,65</td>
<td>95,71</td>
<td>95,97</td>
</tr>
<tr>
<td>No-En</td>
<td>75,71</td>
<td>81,34</td>
<td>86,21</td>
<td>92,28</td>
<td>96,07</td>
</tr>
<tr>
<td>Enh</td>
<td>88,92</td>
<td>89,01</td>
<td>91,54</td>
<td>94,14</td>
<td>96,07</td>
</tr>
<tr>
<td>No-En</td>
<td>73,51</td>
<td>79,25</td>
<td>84,37</td>
<td>90,34</td>
<td>95,54</td>
</tr>
<tr>
<td>Enh</td>
<td>87,23</td>
<td>88,52</td>
<td>89,42</td>
<td>93,25</td>
<td>96,07</td>
</tr>
<tr>
<td>No-En</td>
<td>70,45</td>
<td>77,83</td>
<td>81,26</td>
<td>88,25</td>
<td>94,16</td>
</tr>
<tr>
<td>Enh</td>
<td>86,01</td>
<td>87,64</td>
<td>88,01</td>
<td>89,11</td>
<td>96,07</td>
</tr>
<tr>
<td>No-En</td>
<td>67,33</td>
<td>73,25</td>
<td>79,21</td>
<td>84,07</td>
<td>94,16</td>
</tr>
<tr>
<td>Enh</td>
<td>84,20</td>
<td>87,64</td>
<td>87,25</td>
<td>89,93</td>
<td>96,07</td>
</tr>
<tr>
<td>No-En</td>
<td>65,71</td>
<td>69,42</td>
<td>77,21</td>
<td>80,16</td>
<td>93,22</td>
</tr>
<tr>
<td>Enh</td>
<td>81,52</td>
<td>87,64</td>
<td>88,27</td>
<td>90,93</td>
<td>96,07</td>
</tr>
</tbody>
</table>

Table 1: (%) word recognition performance of degraded/enhanced signals for degradation of up to 90% of the signal waveform at SINRs ranging from -10 to 20 dB.