SPEECH ENHANCEMENT IN WIRELESS DIGITAL COMMUNICATION VIA HEURISTIC RULES AND IMAGE RELAXATION TECHNIQUES

Enzo Mumolo‡ *Lorenzo Pivetta*‡ *Claudio Chiaruttini*†
‡ DEEI, Dipartimento di Elettrotecnica, Elettronica ed Informatica
Università di Trieste, via Valerio 10, 34127 Trieste, Italy
E-mail: mumolo@univ.trieste.it
† Dipartimento di Architettura Navale, Mare, Ambiente
Università di Trieste, via Valerio 10, 34127 Trieste, Italy

ABSTRACT

A novel algorithm for the reduction of several types of noise that occur typically in wireless digital speech communications is described in this paper. The algorithm aims at reducing the spectral discontinuities of the signal by analyzing the 2D spectral map and closing the gaps between the frames using heuristic rules. Some experimental evaluations are reported.

1 Introduction

Wireless digital voice communication suffers of several types of noise. Besides white additive noise, wireless digital communications may be corrupted by micro-interruptions of the signal itself, due to packet transmission errors. Similar problems may also arise in internet telephone transmissions. Analog wireless transmissions, on the other hand, may be corrupted by tonal noise disturbances, due to interference effects.

In this paper we describe an algorithm for reducing such types of noise in voice communications. A common effect of those types of noise is that they introduce an irregularity in the spectrum. The idea of the algorithm is to reconstruct the spectral continuity in the voiced sections of the signal.

We carried out the study with the categories reported above, namely additive white noise, micro-interruptions and bursts of tonal noise. Due to the fact that such noises have similar effects in terms of the above mentioned spectral discontinuities, the algorithm is able to work equally well with the described types of noise without modification.

The algorithm described in this paper uses spectral geometric information and a set of heuristic rules to enhance the speech corrupted by noise. A novel approach for the extraction of those geometric semantic attributes from the speech spectrum, based on texture analysis and scene labeling techniques, is described .

The approach is based on a 2D spectral representation of the input signal. The generic pixel represents the (frequency, frame number) couple of a spectral peak obtained from the Short Time Fourier Transform (STFT) of a given frame of the signal. The phase information accompany the peaks but it is treated differently.

The peaks are less perturbed by noise then other sections of the spectrum, because the noise affects mainly the low energy sections. Thus, spectral peaks are good candidates for the treatment of noisy speech. Moreover, frequencies, phases and amplitudes of spectral peaks are the optimum parameters, in the least squares sense, of the sinusoidal model.

This paper is organized as follows. In section 2 a brief summary of other speech enhancement algorithms developed in the past is reported. Section 3 deals with the sinusoidal model which is the basis of the reconstruction part of the enhancement algorithm. The algorithm is described with some details in section 4, while in section 5 some experimental results of the algorithm on corrupted Italian sentences are reported. The concluding remarks are described in section 6.

2 Related Work

Much work has been developed to the speech enhancement problem. The solutions developed so far can be roughly divided into single filtering, model based and signal periodicities approaches. Let us start with the first approach. A number of methods based on the Fourier Transformation have been classified as spectral subtraction approaches [1]. In other work, an all-pole model of speech is computed and such representation is used to form a Wiener filter which can be iteratively used for reducting the noise level in the signal [2]. Other approaches are based on the minimum mean squared estimation of the conditional probability of the clean signal parameters given the noise observations [5]. In general, the methods based on signal filtering have several drawbacks, such as some residual noise and, for the spectral subtraction method, a great sensitivity with the signal energy. Moreover, if the low energy sections of the signal are attenuated, the intelligibility is reduced.

The signal periodicity methods are based on the idea that the useful signal is essentially related to the high energy portions of the spectrum. Thus, using comb filters [3], these sections are extracted and the signal is

reconstructed.

Finally, other speech enhancement are based on models. The models, for example, can be based on the extracted pitch [4] or on the sinusoidal model [6].

3 The Sinusoidal Model

The sinusoidal model of the speech signal s(n) is described as follows:

$$\hat{s}(n) = Re\left\{\sum_{l=1}^{L_k} A_l^k e^{j[\omega_l^k n + \phi_l^k]}\right\} \qquad (1)$$

where A_l^k, ω_l^k, ϕ_l^k are the sine wave parameters, namely amplitude, frequency and phase of the l-th sine wave for the k-th frame.

The correct reconstruction of the signal, however, requires the matching between adjacent frames peaks. McAulay and Quatieri have proposed a frequency tracker using a birth-death frequency concept [7]. However, this frequency tracker is no longer sufficient to make the correct decisions about matching when the input signal is highly corrupted. By effect of noise, in fact, the peaks can be cancelled, spuriously introduced or moved in frequency. Moreover, gaps in the spectral image can occur. In vowel-like sections of speech, however, the trajectories of the spectral peaks should be highly regular. Such sections, moreover, are very important from a perceptual point of view. The basic idea of our approach, therefore, is to exploit such regularities by looking at the overall behavior of the spectral peaks trajectories rather than at just two adjacent frames.

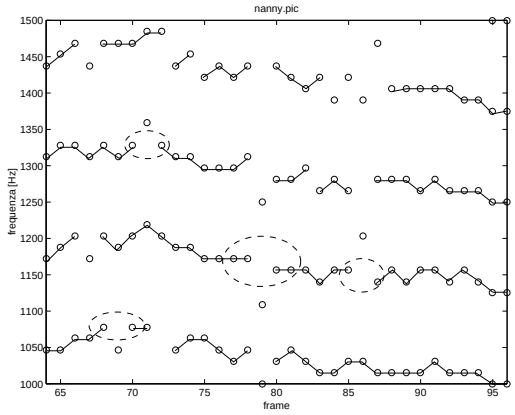

Figure 1: Frames modified by perturbation

4 The Speech Enhancement Algorithm

As mentioned in the previous section, a major problem in the frequency matching process proposed by Quatieri and McAulay is that the matching is performed on pairs of frames only. In many practical cases, however, the perturbations modify several frames of the signal as represented in Fig.1. Thus, two frames are no sufficient for performing successfully the frequency matching. Our algorithm analyzes the matrix of the spectral peaks in order to identify gaps and irregularities in the matrix of spectral peaks and to introduce suitable peaks for restoring the continuity of the trajectories. In Fig.2 the whole process is outlined.

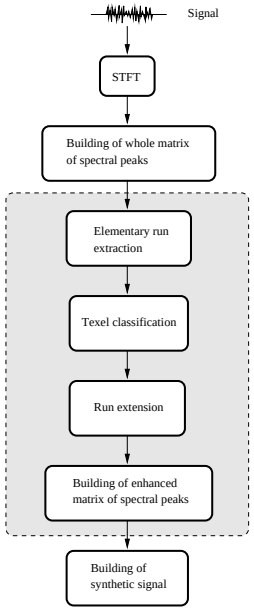

Figure 2: Block diagram of the speech enhancement algorithm

4.1 Extraction of the Elementary Runs

An elementary run is a set of connected points in the spectral matrix. Their extraction is performed on 16x16 pixels texels in which the spectral image is divided. The search for the elementary runs is performed in three directions, as shown in the following pseudocode:

```
For all the pixels in the texel Do
    For direction in 1st_diagonal 2nd_diagonal horizontal Do
        If direction=1st_diagonal Then
            If pixels connected Then save the diagonal run
        If direction=2nd_diagonal Then
            If pixels connected and not belonging to 1st_diagonal
                Then
                save the diagonal run
        If direction=horizontal Then
            If pixels connected and not belonging to 1st, 2nd diag.
                Then
                save the horizontal run
```

To every pixel of the map two bidirectional queues are associated, one describing the pixels and the other one describing the right and left sections of the runs.

4.2 Texel classification

The texels are then classified using the characteristics of the processed runs. Each texel T_{ij} is associated two

measures [9], given by (2) and (3). The first measure is related to the presence of long elementary runs, and the second describes the texels characterized by a small spatial coherence.

$$RE_{ij} = \frac{\sum_{l=1}^{n} l^2 R(l)}{\sum_{l=1}^{n} R(l)} \quad (2)$$

$$UE_{ij} = \frac{\sum_{l=1}^{n} R^2(l)}{\sum_{l=1}^{n} R(l)} \quad (3)$$

Using the described measure, a set of additional measures are computed, aiming at classifying the texels on the basis of the predominant directions of the runs. The additional measures are described by (3) and (4).

$$P_{ij}(k) = max\{0, \overline{RE}_{ij}(k) - \alpha\overline{UE}_{ij}(k)\}, k = 1...7 \quad (4)$$

$$P_{ij}(8) = min\{\overline{UE}_{ij}(k)\}, k = 1...7 \quad (5)$$

The first seven quantities describe the texels on the basis of the main directions of the runs present in the texel itself. The last class characterizes the texels with small spatial coherence.

An additional refinement of the classification is made using the image relaxation technique reported in [8]. Such technique uses the iterative application of the following relations:

$$P_{ij}^{(r+1)} = \frac{P_{ij}^{(r)}[1 + Q_{ij}^{(r)}(k)]}{\sum_l P_{ij}^{(r)}(l)[1 + Q_{ij}^{(r)}(l)]} \quad (6)$$

where the correction factors $Q_{i,j}$ introduce the correlation between the four adjacent texels.

Thus, the generic texel $T_{i,j}$ is classified using the following eight categories:

(*positive_transition, medium_positive_transition, slight_positive_transition, stationary, slight_negative_transition, medium_negative_transition, negative_transition, not_interpreted*)

which correspond to slopes in the peaks trajectories ranging from 45 to -45 degree. Numbering those classes from 1 to 8, the classification is performed as follows:

$$T_{i,j} = argmax\{P_{ij}(k)\}, k = 1...8 \quad (7)$$

It is worth noting that the elementary runs are classified into the above mentioned categories as they are collected.

4.3 Run Extension

The next step of the algorithm is to connect the elementary runs, correcting at the same time the noise effects. This means that the runs interruptions, due to peaks movement and deletion, must be recovered. The connection is performed only in the texels which are classified as belonging to the classes 1 to 7, that is in the texels which are interpreted in a phonetic meaningful way. Moreover, a subset of classified texels is considered, that is the texels corresponding to frequencies below a given threshold. In such texels, a set of heuristic rules for the runs connection are applied. Basically the rules avoid that a cuspid be generated in the connected run, and moreover impose a limit in the slopes of the runs being connected. The heuristic rules for connecting the runs are motivated by phonetic criteria, and are described as follows:

1. The runs are connected according to a minimum distance criterion
2. For phonetically meaningful reasons, it must be avoided the connection between two runs with opposite directions
3. Finally, in case more than one run are available for connection, the closest one has to be chosen

The heuristic rules have been implemented using some search windows, tuned to the three cases of horizontal, diagonal and anti-diagonal (i.e. with negative slope) runs. In fig.3 the search windows related to a connection of diagonal runs with a positive slope are represented as an example.

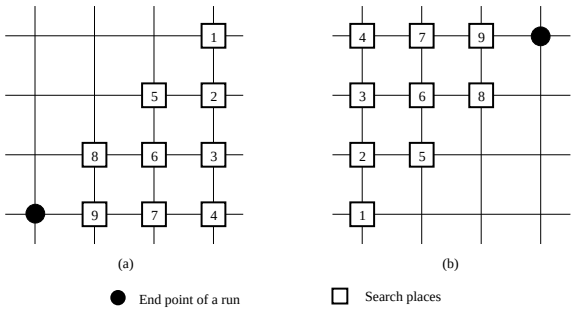

Figure 3: Example search windows

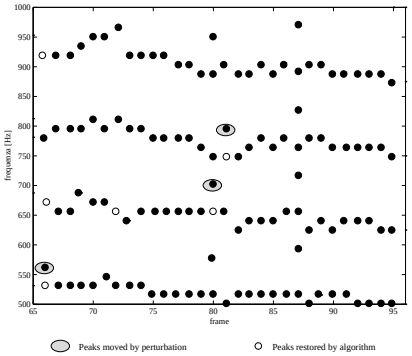

Figure 4: Correction examples

5 Experimental Results

The speech enhancement algorithm has been tested with the types of noise and perturbations which cover

the noise in a mobile communication environment as mentioned above, namely additive white noise, micro-interruptions and sinusoidal bursts. The last two were driven by an exponential distribution of arrivals. Figure 4 shows the corrections performed on voice corrupted by micro-interruptions. Extensive subjective tests were carried out to quantify the performance. In figure 5, the differential MOS improvement over the corrupted voice corresponding to micro-interruption disturbance is reported.

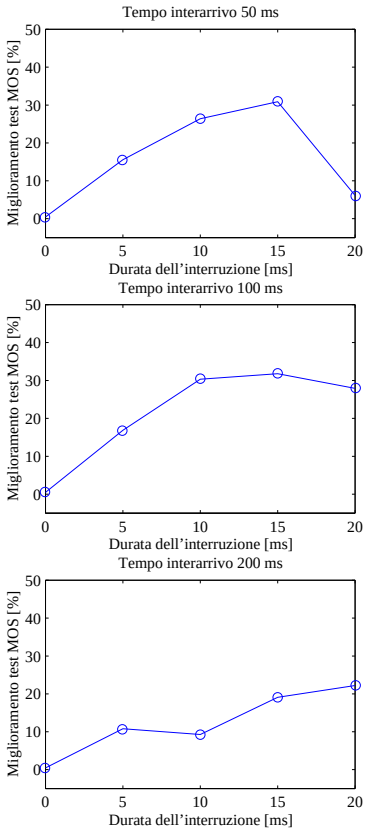

Figure 5: Percentual MOS improvement

In fig.6, moreover, the absolute MOS map describing the subjective quality improvement of utterances degraded with micro-interruptions is reported. The data of figg. 5 and 6 is averaged over different phrases and twenty listeners.

6 Conclusion

In this paper we have described an algorithm based on heuristic rules for the enhancement of speech signals corrupted by several types of noises, such as micro-interruptions, tonal noises and gaussian noise. Some representative results have been reported in case of micro-interruptions, which are similar to the other types of noises. The algorithm requires a quite low computational complexity and it could be used as a preprocessor in digital speech communication environments both for

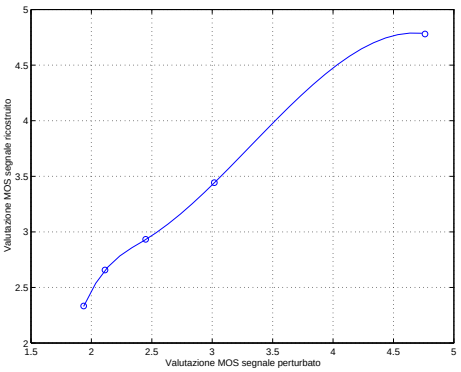

Figure 6: Absolute MOS improvement

improving the coding quality and for improving recognition performances. The algorithm, in the form presented here, is inherently suitable for off-line processing, and an adaptive version is currently underway.

References

[1] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. on ASSP, ASSP-27, 1979

[2] J.S.Lim, A.V.Oppenheim, "All-pole modeling of degraded speech", IEEE Trans. on ASSP, Vol. ASSP-26, 1978

[3] J.S.Lim, A.V.Oppenheim, L.D.Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition", IEEE Trans. on ASSP, Vol.ASSP-26, N.4, 1978

[4] J.D.Wise, J.R.Caprio, T.W.Parks, "Maximum Likelihood pitch estimation", IEEE Trans. on ASSP, Vol.ASSP-24, 1976

[5] Y.Ephraim, D.Malah, "Speech enhancement using a minimum mean-squared error short-time spectral amplitude estimator", IEEE Trans. on ASSP, Vol. ASSP-32, 1984

[6] T.F.Quatieri, R.J.McAulay, "Noise reduction using a soft decision sinewave vector quantizer", ICASSP, Albuquerque, April 90

[7] R.J.McAulay, T.F.Quatieri, "Low Rate Speech Coding Based on the Sinusoidal Coding", in *Advances in Speech Signal Processing*, Marcel Dekker Inc., New York, 1992

[8] R.M.Haralick, "Statistical and Structural Approaches to Texture", *Proceedings of the IEEE*, Vol.67, No.5, May 1979

[9] C.Chiaruttini, P.L.Bragato, G.Cassiani, "Seismic Reflection Interpretation as an Image Understanding Problem", *Journal of Seismic Exploration*, 3, 1994