Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 21 (2007) 174–186
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Iterative reconstruction of speech from short-time Fourier transform phase and magnitude spectra

Leigh D. Alsteris, Kuldip K. Paliwal *

School of Microelectronic Engineering, Griffith University, Brisbane, Qld 4111, Australia

Received 6 May 2005; received in revised form 10 January 2006; accepted 10 March 2006
Available online 6 April 2006

Abstract

In this paper, we consider the topic of iterative, one dimensional, signal reconstruction (specifically speech signals) from the magnitude spectrum and the phase spectrum. While this topic has been extensively researched and documented, we wish to recast some well-established results for the benefit of new researchers and those who desire a short, yet comprehensive, review of the subject. The three main points of the review are: (i) a signal can be reconstructed to within a scale factor from its phase spectrum, (ii) a signal cannot be reconstructed to within a scale factor from its magnitude spectrum, and (iii) a signal can be reconstructed to within a scale factor from its magnitude spectrum when the phase-sign (i.e., one bit of phase spectrum information) is known. Through a number of illustrative examples, we first demonstrate how the algorithms work when the spectral information is determined over the entire duration of the signal. We then demonstrate that the algorithms are equally valid for reconstruction of a signal from the spectra obtained from short-time segments. In addition, we present the results of some further experimentation in which we have attempted to reconstruct a speech signal from only partial phase spectrum information (in the absence of all magnitude spectrum information). We make the following observations: (i) intelligible signal reconstruction (albeit noisy) is possible from knowledge of only the phase spectrum sign information, (ii) an intelligible signal cannot be reconstructed from knowledge of only the phase spectrum frequency-derivative or only the phase spectrum time-derivative, and (iii) an intelligible signal can be reconstructed from the combined knowledge of both the phase spectrum frequency-derivative and time-derivative.

1. Introduction

In automatic speech recognition (ASR), the speech is processed frame-wise using a temporal window duration of 20–40 ms. The short-time Fourier transform (STFT) is normally used for the signal analysis of each frame. The resulting signal spectrum can be decomposed into the magnitude spectrum and the phase spectrum.[1] At such small temporal window durations, it is generally believed that the phase spectrum does not contribute much to speech intelligibility (Liu et al., 1997; Oppenheim and Lim, 1981; Schroeder, 1975) and,

* Corresponding author. Tel.: +61 7 3875 6536; fax: +61 7 3875 5384.
E-mail addresses: L.Alsteris@griffith.edu.au (L.D. Alsteris), K.Paliwal@griffith.edu.au (K.K. Paliwal).

[1] Throughout this paper, the modifier 'short-time' is implied when mentioning the magnitude spectrum and the phase spectrum.

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2006.03.001

as a result, state-of-the-art ASR systems generally discard the phase spectrum in favour of features that are derived only from the magnitude spectrum (Picone, 1993).

We have recently published a number of papers with an intent to provoke fellow researchers to investigate the phase spectrum for use in ASR. This paper also serves the same purpose. Our motivation stems from the results we obtained from some human listening tests (Paliwal and Alsteris, 2003, 2005; Alsteris and Paliwal, 2004); the results indicate that the phase spectrum can contribute significantly to speech intelligibility over small window durations (i.e., 20–40 ms). This is an interesting result, indicating the possible usefulness of the phase spectrum for ASR.

Although the phase spectrum has yet to be proven useful for ASR,[2] it has successfully been used for many other tasks, such as formant extraction (Murthy et al., 1989; Duncan et al., 1989; Potamianos and Maragos, 1996; Friedman, 1985), pitch extraction (Smits and Yegnanarayana, 1995; Satyanarayana and Yegnanarayana, 1999; Abe et al., 1995; Charpentier, 1986; Nakatani et al., 2003), and iterative signal reconstruction (Oppenheim and Lim, 1981; Hayes et al., 1980; Quatieri and Oppenheim, 1981; Tom et al., 1981; Van Hove et al., 1983; Nawab et al., 1983; Merchant and Parks, 1983; Griffin and Lim, 1984; Yegnanarayana et al., 1984, 1987). In this paper, we concern ourselves with iterative signal reconstruction. Any researcher with an interest in the phase spectrum should be aware of the flurry of activity that occurred in the 1980's in the area of signal reconstruction from magnitude spectrum and phase spectrum. In fact, what we find most interesting is that the phase spectrum alone can be used for perfect signal reconstruction (to within a scale factor), yet it has not been used successfully for ASR.

First and foremost, this paper serves as a tutorial on the topic of iterative, one dimensional,[3] signal reconstruction (specifically speech signals). Secondly, we provide the results of some further experimentation which may be interesting from an ASR viewpoint. While iterative signal reconstruction has been extensively researched and documented (Oppenheim and Lim, 1981; Hayes et al., 1980; Quatieri and Oppenheim, 1981; Tom et al., 1981; Van Hove et al., 1983; Nawab et al., 1983; Merchant and Parks, 1983; Griffin and Lim, 1984; Yegnanarayana et al., 1984, 1987), we wish to recast some well-established results for the benefit of new researchers and those who desire a short, yet comprehensive, review of the subject. We believe that an appreciation for how the phase spectrum has proven useful in iterative signal reconstruction will motivate readers to investigate the potential for its use in ASR.

In general, the magnitude and phase spectra are both required in order to uniquely specify a signal. Under certain conditions, however, one can establish relationships between the magnitude and phase spectrum components. A well known result is the relationship of log magnitude spectrum and phase spectrum through the Hilbert transform for minimum and maximum-phase signals (Quatieri and Oppenheim, 1981; Yegnanarayana et al., 1984; Oppenheim and Schafer, 1975). However, finite duration speech signals are mixed-phase, all-zero signals. Hayes et al. (1980) have determined the conditions under which such signals can be uniquely specified to within a scale factor by the phase spectrum, while Van Hove et al. (1983) have determined that such signals can be uniquely specified by the signed-magnitude spectrum (magnitude spectrum with one bit of phase spectrum information). Given the phase spectrum, or signed-magnitude spectrum, the iterative framework in Fig. 1 can be used to reconstruct the signal (where the known spectral information is determined over the entire duration of the signal). This algorithm is equally valid for reconstruction of a signal from short-time segments (Fig. 3).

In this paper, we provide several examples which demonstrate the application of these established iterative signal reconstruction algorithms. We also wish to draw attention to the results of some additional experimentation – since our interest lies in the phase spectrum, we look further into signal reconstruction from the phase spectrum, specifically partial phase spectrum information.[4] The train of thought is that if a signal can be

[2] There have been some attempts at using the phase spectrum as a representation for ASR feature extraction (Murthy and Gadde, 2003; Hegde et al., 2004a,b; Alsteris and Paliwal, 2005; Potamianos and Maragos, 2001; Dimitriadis and Maragos, 2003; Paliwal and Atal, 2003; Wang et al., 2003).

[3] The theory of one dimensional signal reconstruction can be extended to multidimensional signal reconstruction (Oppenheim and Lim, 1981; Hayes et al., 1980; Hayes, 1982).

[4] This work is different to the partial phase spectrum experiments conducted by Yegnanarayana and his colleagues (1987). See Section 3 for more details.

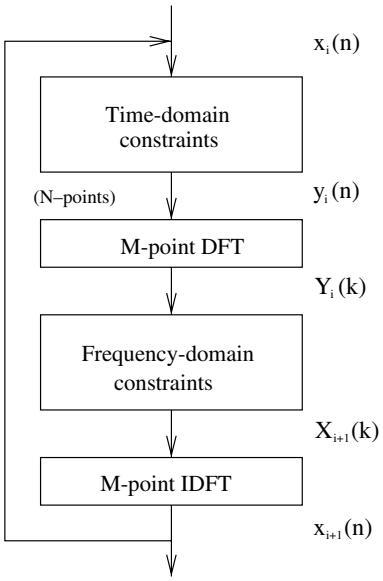

Fig. 1. Iterative framework used for reconstruction of an *N*-point sequence from phase spectrum, magnitude spectrum or signed-magnitude spectrum.

reconstructed from knowledge of only the phase spectrum, why then is the phase spectrum useless for extracting ASR features? If so much information is contained in the phase spectrum, then it may be possible to capture and use it to improve the performance of ASR systems. However, using the phase spectrum directly for ASR has proven difficult due to phase wrapping and other problems (Murthy et al., 1989; Duncan et al., 1989; Yegnanarayana and Murthy, 1992). Here we consider some alternative representations of the phase spectrum. The phase spectrum has two independent variables: frequency and time. Thus, while there may be many ways to represent the information present in the phase spectrum, two representations that first come to mind are those that can be obtained either by taking its frequency-derivative (group delay function, GDF) or its time-derivative (instantaneous frequency distribution, IFD). We want to determine if an intelligible signal can be reconstructed given that we only know either the GDF or the IFD information. We also want to determine if we can reconstruct an intelligible signal given that we only know the phase spectrum sign information. The justification for this further experimentation is as follows: if the use of either the phase spectrum sign, GDF, or IFD information results in intelligible signal reconstruction, this would advocate the possible use of the partial information as a basis for an ASR feature set. Note that there may well be other phase spectrum representations that our readers can investigate.

The paper outline is as follows: In Section 2, we review some established iterative algorithms that attempt to reconstruct a signal from phase spectrum, magnitude spectrum or signed-magnitude spectrum information (where the spectrum is determined over the entire duration of the signal or on a short-time basis). We highlight the fact that knowledge of the phase spectra is enough for unique signal reconstruction (to within a scale factor), while the same is not true for magnitude spectra (the magnitude spectra must be accompanied by phase spectrum sign information for unique reconstruction). In Section 3, we explore the use of partial phase spectrum information, in the absence of all magnitude spectrum information, for intelligible signal reconstruction.

2. An overview of iterative reconstruction algorithms

In this section, we review some well-established signal reconstruction algorithms. We see that under mild conditions, a finite duration signal can be reconstructed to within a scale factor by its phase spectrum (where the phase spectrum is determined over the duration of the signal or on a short-time basis). This is not true for the magnitude spectrum. However, if the magnitude spectrum is accompanied by some phase spectrum information, then unique reconstruction is possible.

2.1. Reconstruction from partial Fourier transform information

2.1.1. Reconstruction from phase spectrum

In practical terms, the theorem proposed by Hayes et al. (1980) (1-d case) is stated as follows:

Theorem 1. *A sequence which is known to be zero outside the interval* $0 \leqslant n \leqslant (N-1)$ *is uniquely specified to within a scale factor by* $(N-1)$ *distinct samples of its phase spectrum in the interval* $0 < w < \pi$ *if it has a z-transform with no zeros on the unit circle or in conjugate reciprocal pairs* (Hayes et al., 1980).

The reconstruction procedure is based on the iterative framework in Fig. 1. In the time domain, all samples outside of the interval $0 \leqslant n \leqslant (N-1)$ are set to zero (i.e., finite-time constraint). In the frequency domain, the known phase spectrum samples are imposed. In order to obtain $N-1$ distinct phase spectrum samples in the interval $0 < w < \pi$, a discrete Fourier transform (DFT) of length $M \geqslant 2N$ is required. In our experiments, we use a DFT length of $M = 2N$. Repeated transformations between the time and frequency domains, with the continued enforcement of the above constraints, provide a signal that converges to a scaled version of the original signal (Tom et al., 1981).

This algorithm has been used to reconstruct the signal in Fig. 2(a) from its phase spectrum. The magnitude spectrum is initially set to unity for all ω. The reconstructed signal after 200 iterations is shown in Fig. 2(b).

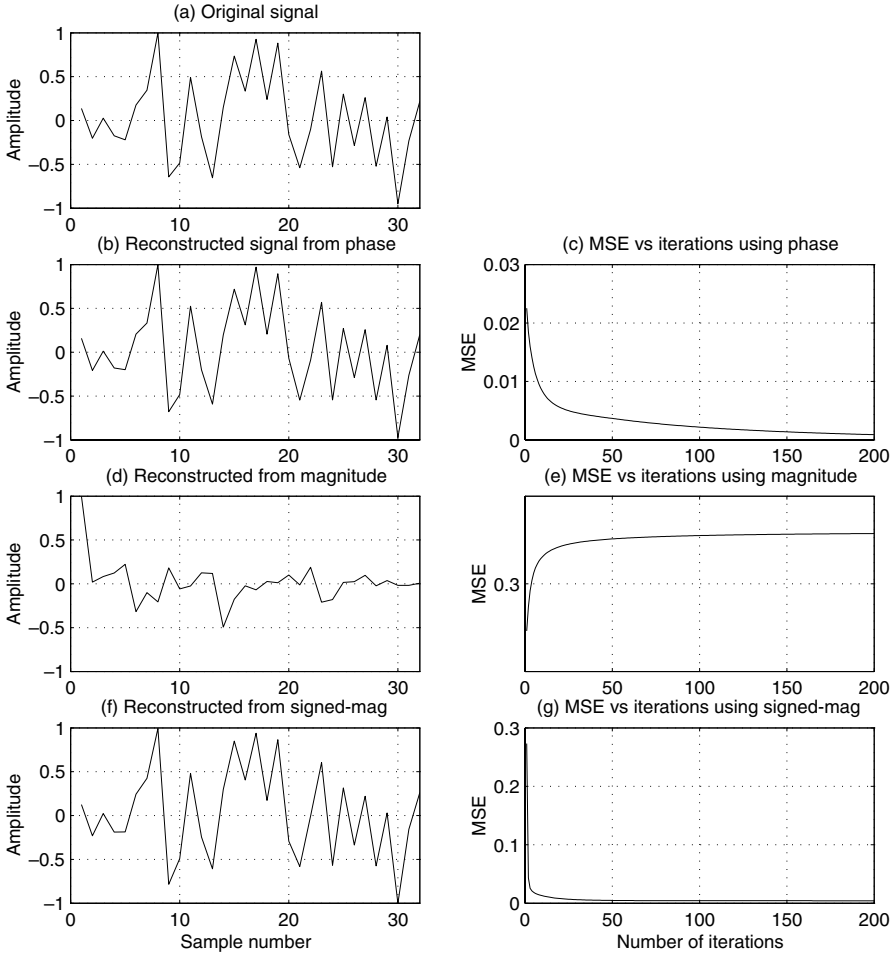

Fig. 2. Results of experiments in Section 2.1. (a) is the original signal, (b), (d) and (f) show reconstructed signals after 200 iterations (these signals are scaled to vary over the same range as the original signal). (c), (e) and (g) show plots of the respective MSE values at every iteration.

The mean squared error (MSE) between the original and reconstructed signals[5] is non-increasing with each iteration (Fig. 2(c)).

2.1.2. Reconstruction from magnitude spectrum

Unlike the phase spectrum, the magnitude spectrum cannot uniquely specify a 1-d sequence. The reason is as follows. If we express the z-transform of the signal as:

$$S(z) = G\prod_{k=1}^{M}(1 - b_k z^{-1}), \tag{1}$$

where G is real, then the square of the magnitude function is expressed as (Oppenheim and Schafer, 1975):

$$P(z) = S(z)S^*(1/z^*) = G^2\prod_{k=1}^{M}(1 - b_k z^{-1})(1 - b_k^* z). \tag{2}$$

The zeros occur in conjugate reciprocal pairs. Thus the zeros of $S(z)$ cannot be determined by the magnitude spectrum alone. Therefore, the phase spectrum can not be determined by the magnitude spectrum. Consequently, if the known magnitude spectrum (instead of the phase spectrum) is imposed in the iterative reconstruction algorithm, it will not converge to the original signal.

If the signal is assumed to be minimum or maximum phase, then there is no ambiguity in determining the zeros from magnitude.[6] In the case of mixed-phase signals, Van Hove et al. (1983) have shown that this ambiguity can be resolved by imposing some phase spectrum information (see Section 2.1.3).

We reconstruct the signal in Fig. 2(a) from its magnitude spectrum. The phase spectrum is initialised with random values. After 200 iterations, the reconstructed signal does not resemble the original signal (Figs. 2(d) and (e)).

2.1.3. Reconstruction from signed-magnitude spectrum

The Fourier transform phase spectrum is defined by:

$$\phi(\omega) = \arctan(S_i(\omega)/S_r(\omega)) \tag{3}$$

where the arctangent provides values in the range $[-\pi, \pi]$. Therefore, included in the knowledge of $\phi(\omega)$ are the signs of the real and imaginary components. Van Hove et al. (1983) show that the magnitude spectrum, along with this sign information, provides a unique specification of a finite duration causal sequence. The 'signed-magnitude' is defined as,

$$A(\omega : \omega_0) = \begin{cases} |S(\omega)| & \text{if } -\omega_0 \leqslant \phi(\omega) < \omega_0 + \pi, \\ -|S(\omega)| & \text{otherwise,} \end{cases} \tag{4}$$

where ω_0 is an arbitrary number within the interval $[-\pi, \pi)$. Thus, $A(\omega{:}\omega_0)$ contains information about both the magnitude spectrum and the sign of the real and imaginary parts of the Fourier transform. Their theorem is stated as follows:

Theorem 2. *Let $x(n)$ and $y(n)$ be two real, causal, and finite extent sequences with z-transforms which have no zeros on the unit circle. If $A_x(\omega{:}\omega_0) = A_y(\omega{:}\omega_0)$ for all ω then $x(n) = y(n)$* (Van Hove et al., 1983).

In terms of the iterative reconstruction algorithm, $A(\omega{:}\omega_0)$ imposes both a magnitude spectrum and phase spectrum constraint. When both of these constraints are enforced, the algorithm converges to the original signal (Figs. 2(f) and (g)). In our experiments, we use $\omega_0 = \pi/2$.

The phase spectrum constraint amounts to splitting the phase spectrum in half (at an arbitrary point), then taking note in which half the phase spectrum values lie for each frequency. In every iteration of the reconstruction algorithm, each phase spectrum value is constrained to vary only within the half from which the original phase spectrum value came. This is enforced by adding π to any phase values that are not in the correct half.

[5] In all experiments, the signals reconstructed from phase spectrum are rescaled to vary over the same range as the original signal. MSE measurements are taken after rescaling.

[6] Reconstruction algorithms for minimum phase signals can be found in Quatieri and Oppenheim (1981), Yegnanarayana et al. (1984).

2.2. Reconstruction within the STFT framework

Theorems 1 and 2 are also applicable in the context of the STFT (Yegnanarayana et al., 1987). The STFT overlap-add analysis imposes additional restrictions on magnitude-phase pairings. Specifically, adjacent short-time sections must be consistent in their region of overlap. Thus, when reconstructing from partial information, extra information is present in the overlapping sections.

There are two ways to reconstruct via the STFT. One method is referred to as 'sequential extrapolation', where the short-time sections are reconstructed in the order determined by their positions on the time axis. Each section is determined by its known spectral information as well as the known samples in the region of overlap with previous sections. This method is investigated by Nawab et al. (1983). The framework for the method we use is illustrated in Fig. 3. This method was employed by Griffin and Lim (1984) for time-scale modification of speech. It is referred to as 'simultaneous extrapolation'. In this method, the known spectral information of all short-time sections are used simultaneously to determine the unknown signal (i.e., the whole signal is analysed and synthesised in every iteration). In the experiments that follow, we use a rectangular analysis window of duration 32 ms. Any comments made with respect to signal intelligibility are based on informal listening tests by the authors.

2.2.1. Reconstruction from STFT phase spectra

We analyse the signal in Fig. 4(a) at various segment shifts ($\frac{1}{2}, \frac{1}{4}$, and $\frac{1}{8}$), keeping only the phase spectra from each segment. The known phase spectra samples are enforced for every iteration of the reconstruction algorithm. The magnitude spectrum is initially set to unity for all frequencies. For all segment shifts, the algorithm converges toward a scaled version of the original signal (Figs. 4(b) and (c)). The iterative application of overlap and addition ensures that adjacent short-time sections are consistent in their regions of overlap. The greater these regions of overlap, the faster the convergence due to the fact that more initial constraints are imposed on the final solution. This imposition of more constraints also leads to lower MSE. The associated spectrogram is given in Fig. 5(b). It looks identical to the spectrogram of the original signal in Fig. 5(a).

Note that, for the case of reconstruction from short-time phase spectra, there must be at least one sample of overlap between segments. The overlapping sample(s) serve to maintain the energy relationship between

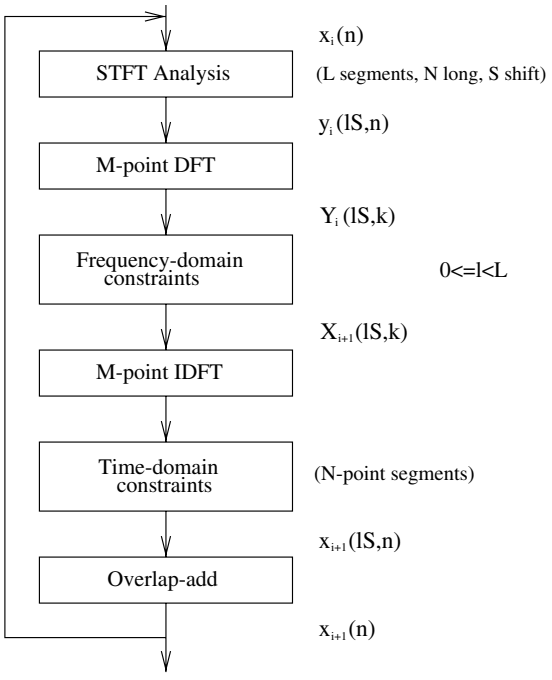

Fig. 3. STFT-based iterative reconstruction framework.

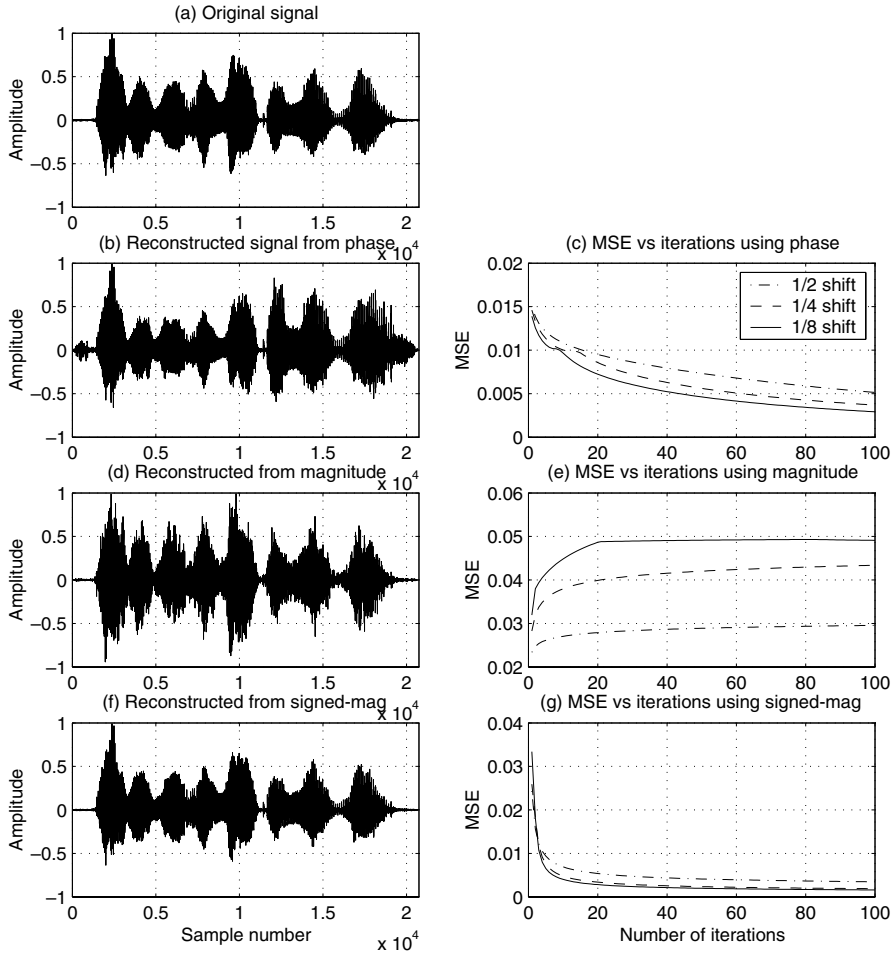

Fig. 4. Results of experiments in Section 2.2. (a) is the original signal used for experiments in Sections 2.2 and 3, "Why were you away a year Roy?". (b), (d) and (f) show reconstructed signals after 100 iterations, using a frame-shift of $\frac{1}{8}$ and a rectangular analysis window of duration 32 ms (these signals are scaled to vary over the same range as the original signal). (c), (e) and (g) show plots of the respective MSE values at every iteration.

adjacent segments. So, even though no energy information is provided to seed the iterative algorithm, the energy contour (albeit scaled) is preserved.

2.2.2. Reconstruction from STFT magnitude spectra

Griffin and Lim first used the STFT reconstruction framework of Fig. 3 to reconstruct time-scaled versions of a signal from short-time magnitude spectra (Griffin and Lim, 1984). Here, we analyse the algorithm for no time-scaling, imposing the known short-time magnitude spectra in each iteration. The short-time phase spectrum for each segment is initially randomised. The algorithm does not converge toward the original speech (Figs. 4(d) and (e)). Informal listening tests, however, indicate that more overlap between frames (i.e., less shift) leads to the reconstructed speech sounding more like the original. This is expected, since more overlap imposes more restrictions on the form of the final solution. The spectrogram of the reconstructed signal is given in Fig. 5(c).

2.2.3. Reconstruction from STFT signed-magnitude spectra

Here, we enforce the known short-time magnitude spectra in addition to the phase spectra sign information (see Section 2.1.3). Once again, we observe the signal converging, with the rate of convergence increasing, and

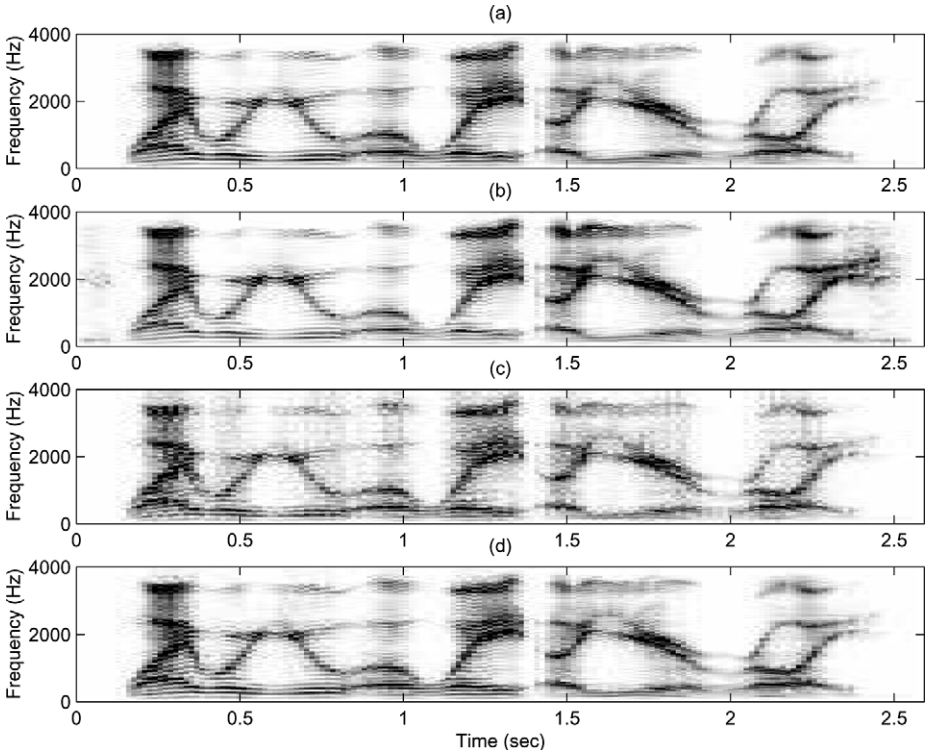

Fig. 5. Spectrograms of (a) original signal "Why were you away a year Roy?", (b) iterative reconstruction from short-time phase spectra, (c) iterative reconstruction from short-time magnitude spectra, (d) iterative reconstruction from short-time signed-magnitude spectra. The signals in (b), (c) and (d) are reconstructed using 100 iterations, a frame-shift of $\frac{1}{8}$ and a rectangular analysis window of duration 32 ms.

the error reducing, with more overlap (Figs. 4(f) and (g)). The spectrogram of the reconstructed signal in Fig. 5(d) looks identical to that of the original signal.

3. Reconstruction from partial STFT phase spectra

In light of results from the previous section (which are well-established), we note that knowledge of the phase spectrum is enough for unique signal reconstruction (to within a scale factor), while the same is not true for magnitude spectrum – the magnitude spectrum must be accompanied by phase spectrum sign information for unique reconstruction. We find it interesting that the phase spectrum can be used to reconstruct a signal, while it is useless for extracting ASR features. If so much information is contained in the phase spectrum, then it may be possible to capture and use it to improve the performance of ASR systems. However, using the phase spectrum directly for ASR has proven difficult due to phase wrapping and other problems (Murthy et al., 1989; Duncan et al., 1989; Yegnanarayana and Murthy, 1992). If there is to be any chance of using the phase spectrum for ASR, then it will have to be via an alternative representation.

To this end, we further analyse the use of the phase spectrum for signal reconstruction. Specifically, we explore the use of partial phase spectrum information (in the absence of all magnitude spectrum information) for intelligible signal reconstruction.[7] At this point, we only hypothesize that it may be possible to capture and use partial phase information to improve the performance of ASR systems. However, we can not theoretically

[7] This work is different to the partial phase spectrum experiments conducted by Yegnanarayana and his colleagues (1987). They use the word 'partial' to denote the situation where the required number of phase spectrum samples for unique signal reconstruction are not known. In order to compensate for the unknown phase spectrum samples, they enforce some known signal samples or magnitude spectrum samples during the iterative reconstruction procedure. In our work, the word 'partial' is used to mean something different. Continue to read Section 3 for an explanation.

justify what exactly that partial information is. To begin with, we can only choose from some well-known partial phase signal representations. There may well be some other representations that our readers can investigate.

We employ the STFT-based reconstruction framework in Fig. 3. The partial phase spectrum representations that we investigate are the phase spectrum sign information, GDF and the IFD. If the use of either the phase spectrum sign, GDF, or IFD information results in intelligible signal reconstruction, this would advocate the possible use of the partial information as a basis for an ASR feature set.

3.1. Reconstruction from STFT phase spectra sign

A similar experiment to that in Section 2.2.3 is performed, however magnitude spectra values are not enforced. Initially, all magnitude spectrum values are set to unity and phase spectrum values are randomised, but given the correct sign. After every iteration, magnitude values are left alone and absolute phase values are left alone. Wherever the phase sign has changed, it is corrected. At first glance, it appears that the MSE does not increase in each iteration (Figs. 6(a) and (b)). However, closer inspection reveals that the MSE increases slightly at some points along the curve. More overlap seems to result in a better estimation of the original signal. Informal listening tests indicate that more overlap also leads to better intelligibility. It is interesting to note that only a small amount of phase spectra information provides for an intelligible signal (although the

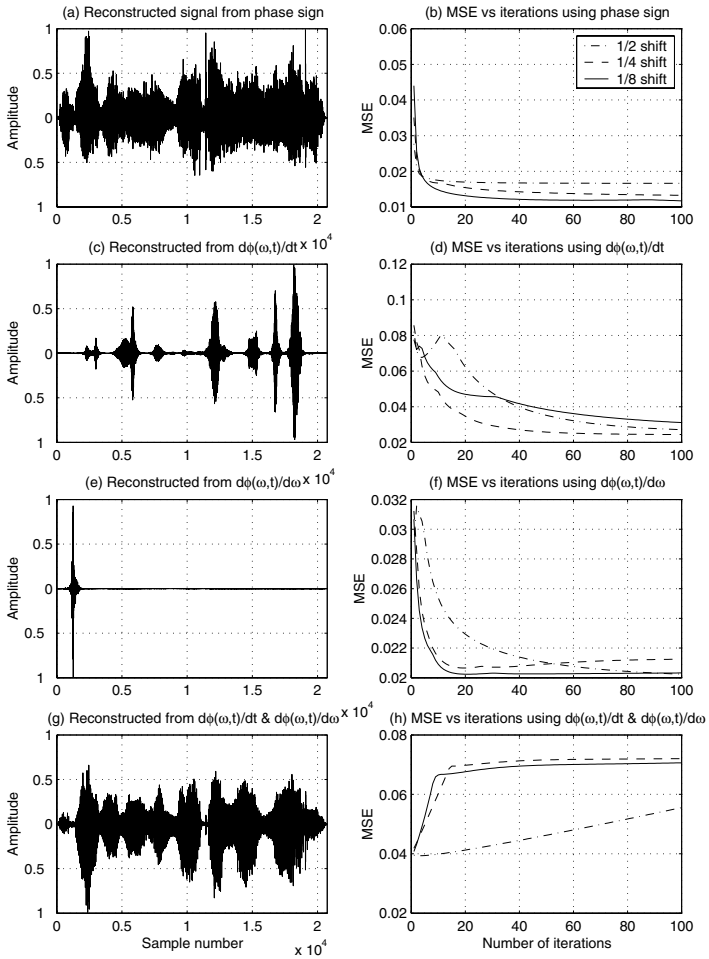

Fig. 6. Results of experiments in Section 3. (a), (c), (e) and (g) show reconstructed signals after 100 iterations, using a frame-shift of $\frac{1}{8}$ and a rectangular analysis window of duration 32 ms (these signals are scaled to vary over the same range as the original signal). (b), (d), (f) and (h) show plots of the respective MSE values at every iteration.

reconstructed signal is noisy). The increased overlap accommodates, to some extent, for the sparse phase spectral information. The spectrogram of the reconstructed signal is given in Fig. 7(a).

3.2. Reconstruction from time and frequency derivatives of STFT phase spectra

We take the phase spectrum from each short-time section and randomise it across frequency, such that the IFD (i.e., $\mathrm{d}\phi(\omega,t)/\mathrm{d}t$) is preserved and the GDF (i.e., $\mathrm{d}\phi(\omega,t)/\mathrm{d}\omega$) is discarded. Specifically, add the same random sequence (across frequency) to the phase spectrum values of each frame. For example, consider a frame of length N and a DFT length of $2N$. Add a random sequence to the phase values in the first $N+1$ DFT bins (i.e., bin numbers 0 to N). To determine the remaining $N-1$ phase values (i.e., bin numbers $N+1$ to $2N-1$), take the new phase values from bins 1 to $N-1$ then reverse the sign and reverse the order of the numbers. That is, given the new phase values for the first $N+1$ bins, calculate the remaining bin phase values by $\phi(n) = -\phi(2N-n)$, where $n = N+1, N+2, \ldots,2N-1$ is the bin number. The resulting IFD-preserved phase spectra are used in place of the original phase spectra in the STFT reconstruction algorithm (and magnitude spectra are initially set to unity). The algorithm does not converge toward the original signal, nor does it provide an intelligible signal (Figs. 6(c) and 7(b)).

In a similar vein, we take the original phase spectra and randomise them across time, such that the GDF is preserved and the IFD is discarded. That is, generate a random sequence whose length is equal to the number of frames in the utterance, then add this same sequence to the time-trajectory of the phase spectrum values for each DFT bin. Remember to do this for the phase values in the first $N+1$ DFT bins (for each frame), then calculate the remaining bin phase values as described above. Reconstruction is performed with the resulting GDF-preserved phase spectra (and magnitude spectra are set to unity). Again, the reconstruction algorithm does not converge to an intelligible solution (Figs. 6(e) and 7(c)).

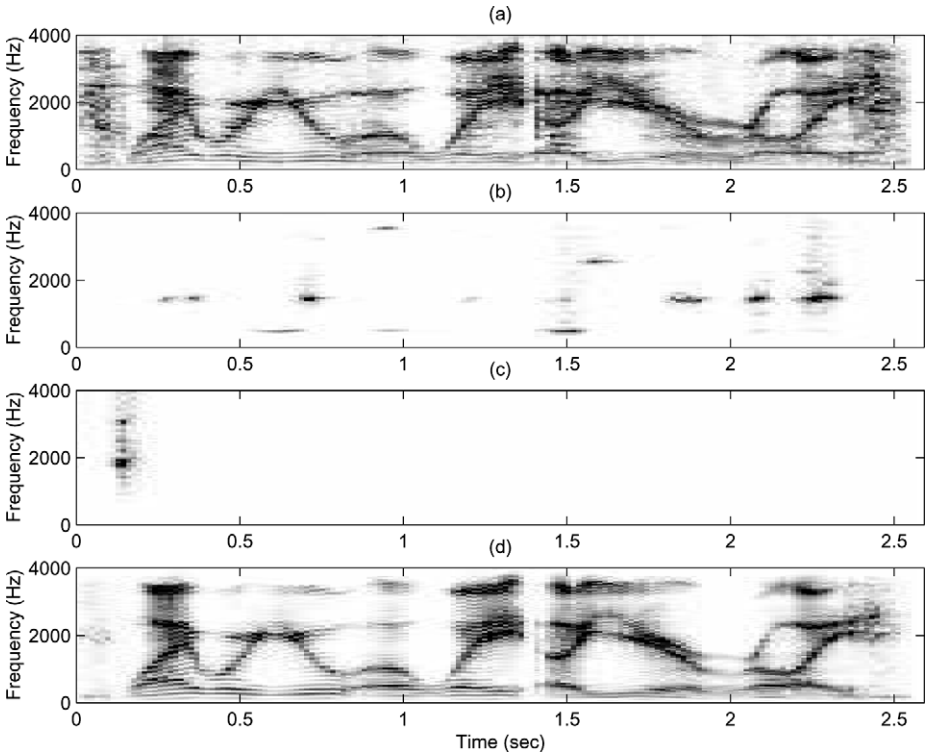

Fig. 7. Spectrograms of reconstructed signals. Signals are reconstructed from knowledge of (a) short-time phase spectrum sign information, (b) $\mathrm{d}\phi(\omega,t)/\mathrm{d}t$, (c) $\mathrm{d}\phi(\omega,t)/\mathrm{d}t$, and (d) both time and frequency derivatives of the short-time phase spectra. All signals are reconstructed using 100 iterations, a frame-shift of $\frac{1}{8}$ and a rectangular analysis window of duration 32 ms. The spectrogram of the original signal is given in Fig. 5(a).

Therefore, reconstruction of intelligible speech is not possible from either knowledge of only the IFD or the GDF. We observed that this holds true regardless of the amount of overlap. Figs. 6(d) and (f) seem to indicate convergence. This is deceiving. In fact, the MSE is converging toward the original signal mean squared amplitude (0.0194), since the algorithm (in both cases) provides a signal whose energy tends to diminish with each iteration.

We now attempt to reconstruct the signal from knowledge of both the GDF-preserved phase spectra and the IFD-preserved phase spectra. In order to do this, we must first extract the GDF information from the GDF-preserved phase spectra and the IFD information from the IFD-preserved phase spectra. Also, notice that the first-segment phase spectrum can only be reconstructed to within a time-shift of the original first-segment phase spectrum, since all we know about it is $\mathrm{d}\phi(\omega,t)/\mathrm{d}t$. The remaining segments are reconstructed in relation to this segment. The details for reconstructing the phase spectrum values from the GDF-preserved phase spectra and the IFD-preserved phase spectra are as follows: First we calculate the GDF for each segment from the GDF-preserved phase spectra. Next, we calculate delta's across time from the IFD preserved phase spectra. The phase value for DFT bin number 0 is set to zero in every frame. The remaining phase values (for each frame) are calculated by cumulatively summing the GDF across DFT bins 1 to N. We then shift all of these phase values by a constant in each frame (dependent on the frame), so that the phase changes over time for that one particular DFT bin (this can be any bin, the decision is arbitrary) are the same as in the original signal (i.e., we use the IFD values for only one bin). The phase values for bins $N+1$ to $2N-1$ are calculated as previously described.[8] Since the original phase spectra values cannot be recovered,[9] the algorithm does not converge (Fig. 6(h)). Regardless of this, a solution that sounds almost exactly like the original speech is provided (Fig. 6(g)). The reconstructed signal is similar to the original signal (Fig. 4(a)) in many respects, apart from the fact that it looks upside-down (which has no effect on intelligibility). The spectrogram in Fig. 7(d) is almost identical to that of the original in Fig. 5(a). Therefore, in the context of the STFT reconstruction framework, when both the IFD and GDF are preserved, adequate information is available for intelligible signal reconstruction.

4. Conclusion

In this paper, we provided a tutorial on the topic of iterative, one dimensional, signal reconstruction (specifically speech signals) from the magnitude spectrum and the phase spectrum. While this topic has been extensively researched and documented, our intention was to recast some well-established results for the benefit of new researchers and those who desire a short, yet comprehensive, review of the subject. The three main points of the tutorial are: (i) a signal can be reconstructed to within a scale factor from its phase spectrum, (ii) a signal cannot be reconstructed to within a scale factor from its magnitude spectrum, and (iii) a signal can be reconstructed to within a scale factor from its magnitude spectrum when the phase-sign (i.e., one bit of phase information) is known. Through a number of illustrate examples, we first demonstrated how the algorithms work when the spectral information is determined over the entire duration of the signal. We then demonstrated that the algorithms are equally valid for reconstruction of a signal from the spectra obtained from short-time segments. In addition, we presented the results of some further experimentation in which we have attempted to reconstruct a speech signal from only partial phase spectrum information (in the absence of all magnitude spectrum information). We make the following observations: (i) intelligible signal reconstruction (albeit noisy) is possible from knowledge of only the phase spectrum sign information, (ii) an intelligible signal cannot be reconstructed from knowledge of only the phase spectrum frequency-derivative or only the phase spectrum time-derivative, and (iii) an intelligible signal can be reconstructed from the combined knowledge of both the phase spectrum frequency-derivative and time-derivative.

[8] Note that this is only one way of reconstructing the phase spectrum values. It is also possible to reconstruct by using the GDF values for only one frame then to extrapolate the phase values for the other frames by using the IFD values for all DFT bins.

[9] The raw phase spectrum values are only meaningful in the context of a fixed-time reference. All that we have lost in this reconstructed signal is the original fixed-time reference. Time referencing is now in relation to the phase spectrum values of the first frame (i.e., we still have a time reference, but it is different to that of the original phase spectra values).

References

Abe, T., Kobayashi, T., Imai, S., 1995. Harmonics tracking and pitch extraction based on instantaneous frequency. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, 756–759.

Alsteris, L.D., Paliwal, K.K., 2004. Importance of window shape for phase-only reconstruction of speech. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing I, 573–576.

Alsteris, L.D., Paliwal, K.K., 2005. Evaluation of the modified group delay feature for isolated word recognition. In: Proceedings of the International Symposium on Signal Processing and Applications, August.

Charpentier, F.J., 1986. Pitch detection using the short-term phase spectrum. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, 113–116.

Dimitriadis, D., Maragos, P., 2003. Robust energy demodulation based on continuous models with application to speech recognition. In: Proceedings of the Eurospeech, September, pp. 2853–2856.

Duncan, G., Yegnanarayana, B., Murthy, Hema A., 1989. A nonparametric method of formant estimation using group delay spectra. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, 572–575.

Friedman, David H., 1985. Instantaneous-frequency distribution vs. time: an interpretation of the phase structure of speech. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, 1121–1124.

Griffin, D.W., Lim, J.S., 1984. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. Speech Signal Processing ASSP-32 (2), 236–243.

Hayes, M.H., 1982. The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform. IEEE Trans. Acoust. Speech Signal Processing ASSP-30 (2), 140–154.

Hayes, M.H., Lim, J.S., Oppenheim, A.V., 1980. Signal reconstruction from phase or magnitude. IEEE Trans. Acoust. Speech Signal Processing ASSP-28 (6), 672–680.

Hegde, R.M., Murthy, H.A., Rao, G.V.R., 2004a. Continuous speech recognition using joint features derived from the modified group delay function and MFCC. In: Proceedings of the International Conference on Speech and Language Processing, October.

Hegde, R.M., Murthy, H.A., Rao, G.V.R., 2004b. The modified group delay feature: a new spectral representation of speech. In: Proceedings of the International Conference on Speech and Language Processing, October.

Liu, L., He, J., Palm, G., 1997. Effects of phase on the perception of intervocalic stop consonants. Speech Communication 22 (4), 403–417.

Merchant, G.A., Parks, T.W., 1983. Reconstruction of signals from phase: efficient algorithms, segmentation, and generalisations. IEEE Trans. Acoust. Speech Signal Processing ASSP-31 (5), 1135–1147.

Murthy, H.A., Gadde, V., 2003. The modified group delay function and its application to phoneme recognition. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing I, 68–71.

Murthy, Hema A., Madhu Murthy, K.V., Yegnanarayana, B., 1989. Formant extraction from Fourier transform phase. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, 484–487.

Nakatani, T., Irino, T., Zolfaghari, P., 2003. Dominance spectrum based v/uv classification and F_o estimation. In: Proceedings of the Eurospeech, September, pp. 2313–2316.

Nawab, S.H., Quatieri, T.F., Lim, J.S., 1983. Signal reconstruction from short-time Fourier transform magnitude. IEEE Trans. Acoust. Speech Signal Processing ASSP-31 (4), 986–998.

Oppenheim, A.V., Lim, J.S., 1981. The importance of phase in signals. Proc. IEEE 69 (May), 529–541.

Oppenheim, A.V., Schafer, R.W., 1975. Digital Signal Processing. Prentice-Hall, Englewood Cliffs, NJ.

Paliwal, K.K., Alsteris, L.D., 2005. On the usefulness of STFT phase spectrum in human listening tests. Speech Communication 45 (2), 153–170.

Paliwal, K.K., Atal, B.S., 2003. Frequency-related representation of speech. In: Proceedings of the Eurospeech, September, pp. 65-68.

Paliwal, K.K., Alsteris, L., 2003. Usefulness of phase spectrum in human speech perception. In: Proceedings of the Eurospeech, Geneva, Switzerland, September, pp. 2117–2120.

Picone, J.W., 1993. Signal modeling techniques in speech recognition. Proc. IEEE 81 (9), 1215–1247.

Potamianos, A., Maragos, P., 1996. Speech formant frequency and bandwidth tracking using multiband energy demodulation. J. Acoust. Soc. Am. 99, 3795–3806.

Potamianos, A., Maragos, P., 2001. Time-frequency distributions for automatic speech recognition. IEEE Trans. Speech Audio Processing 9 (Mar.), 196–200.

Quatieri, J.E., Oppenheim, A.V., 1981. Iterative techniques for minimum phase signal reconstruction from phase or magnitude. IEEE Trans. Acoust. Speech Signal Processing ASSP-29 (6), 1187–1193.

Satyanarayana, P., Yegnanarayana, B., 1999. Robustness of group-delay based method for extraction of significant instants of excitation from speech signals. IEEE Trans. Speech Audio Processing 7 (6), 609–619.

Schroeder, M.R., 1975. Models of hearing. Proc. IEEE 63, 1332–1350.

Smits, R., Yegnanarayana, B., 1995. Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech Audio Processing 3 (5), 325–333.

Tom, V.T., Quatieri, T.F., Hayes, M.H., McClellan, J.H., 1981. Convergence of iterative nonexpansive signal reconstruction algorithms. IEEE Trans. Acoust. Speech Signal Processing ASSP-29 (5), 1052–1058.

Van Hove, P.L., Hayes, M.H., Lim, J.S., Oppenheim, A.V., 1983. Signal reconstruction from signed Fourier transform magnitude. IEEE Trans. Acoust. Speech Signal Processing ASSP-31 (5), 1286–1293.

Wang, Y., Hansen, J., Allu, G.K., Kumaresan, R., 2003. Average instantaneous frequency and average log envelopes for ASR with the aurora 2 database. In: Proceedings of the Eurospeech, September, pp. 25–28.

Yegnanarayana, B., Murthy, H.A., 1992. Significance of group delay functions in spectrum estimation. IEEE Trans. Signal Processing 40 (9), 2281–2289.

Yegnanarayana, B., Saikia, D.K., Krishnan, T.R., 1984. Significance of group delay functions in signal reconstruction from spectral magnitude or phase. IEEE Trans. Acoust. Speech Signal Processing ASSP-32 (3), 610–623.

Yegnanarayana, B., Tanveer Fathima, S., Murthy, H.A., 1987. Reconstruction from Fourier transform phase with applications to speech analysis. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, 301–304.