Some Experiments on Iterative Reconstruction of Speech from STFT Phase and Magnitude Spectra

Leigh D. Alsteris and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University, Brisbane, Australia
L.Alsteris@griffith.edu.au, K.Paliwal@griffith.edu.au

Abstract

In our earlier work, we have measured human intelligibility of stimuli reconstructed either from the short-time magnitude spectra or short-time phase spectra of a speech signal. We demonstrated that, even for small analysis window durations of 20-40 ms (of relevance to automatic speech recognition), the short-time phase spectrum can contribute to speech intelligibility as much as the short-time magnitude spectrum. Reconstruction was performed by overlap-addition of modified short-time segments, where each segment had either the magnitude or the phase of the corresponding original speech segment. In this paper, we employ an iterative framework for signal reconstruction. With this framework, we see that a signal can be reconstructed to within a scale factor when only phase is known, while this is not the case for magnitude. The magnitude must be accompanied by sign information (i.e., one bit of phase information) for unique reconstruction. In the absence of all magnitude information, we explore how much phase information is required for intelligible signal reconstruction. We observe that (i) intelligible signal reconstruction (albeit noisy) is possible from knowledge of only the phase sign information, and (ii) when both time and frequency derivatives of phase are known, adequate information is available for intelligible signal reconstruction. In the absence of either derivative, an unintelligible signal results.

1. Introduction

The Fourier transform of a speech signal $s(t)$ is expressed as,

$$\begin{aligned} S(\omega) &= \int_{-\infty}^{\infty} s(t)e^{-j\omega t}dt \\ &= |S(\omega)|e^{j\phi(\omega)}, \end{aligned} \quad (1)$$

where $|S(\omega)|$ is the Fourier magnitude spectrum and $\phi(\omega) = \angle S(\omega)$ is the Fourier phase spectrum. In general, the magnitude and phase are both required in order to uniquely specify the signal $s(t)$. Under certain conditions, however, one can establish relationships between magnitude and phase components. A well known result is the relationship of log magnitude and phase spectrum through the Hilbert transform for minimum and maximum-phase signals [1, 2, 3]. However, finite speech signals are mixed-phase, all-zero signals. Hayes et al. [4], have determined the conditions under which such signals can be uniquely specified to within a scale factor by phase, while Van Hove et al. [5] have determined that such signals can be uniquely specified by the signed-magnitude (magnitude with one bit of phase information). Accompanying these theorems is an iterative reconstruction framework (Fig. 1), which serves to reconstruct a signal from partial information.

In a recent study [6], we have conducted a number of human perception experiments, the results of which indicate that the phase spectrum can contribute significantly to speech intelligibility over small window durations of 20-40 ms. These results may have direct implications for ASR. Reconstruction was performed by overlap-addition of modified short-time segments, where each segment had either the magnitude or the phase of the corresponding original speech segment. In this paper, we employ the aforementioned iterative framework for signal reconstruction from phase or magnitude. In this framework, we determine to what extent magnitude and phase provide intelligibility.

The paper outline is as follows: In Section 2, we review some well established algorithms that attempt to reconstruct a signal from phase, magnitude or signed-magnitude information (where the spectrum is determined over the entire duration of the signal). In Section 3, we demonstrate that these algorithms are also valid in a short-time framework. We note that knowledge of the phase is enough for unique signal reconstruction (to within a scale factor), while the same is not true for magnitude (magnitude must be accompanied by phase-sign information for unique reconstruction). In Section 4, we explore the use of partial phase information, in the absence of all magnitude information, for intelligible signal reconstruction. Note that intelligibility is determined by informal listening tests by the authors. Quantitative measurements of intelligibility are not provided.

2. Revision of reconstruction algorithms

2.1. Reconstruction from phase

In practical terms, the theorem proposed by Hayes et al. [4] (1-d case) is stated as follows:

Theorem 1 *A sequence which is known to be zero outside the interval* $0 \leq n \leq (N-1)$ *is uniquely specified to within a scale factor by* $(N-1)$ *distinct samples of its phase in the interval* $0 < w < \pi$ *if it has a z-transform with no zeros on the unit circle or in conjugate reciprocal pairs [4].*

The reconstruction procedure is based on the iterative framework in Fig. 1. In the time domain, all samples outside of the interval $0 \leq n \leq (N-1)$ are set to zero (i.e., finite-time constraint). In the frequency domain, the known phase samples are imposed. In order to obtain $N-1$ distinct phase samples in the interval $0 < w < \pi$, a discrete Fourier transform (DFT) of length $M \geq 2N$ 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, provides a signal that converges to a scaled version of the original signal [7].

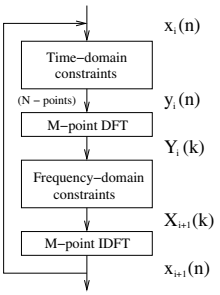

Figure 1: Iterative framework used for reconstruction of an N-point sequence from phase, magnitude or signed-magnitude.

This algorithm has been used to reconstruct the signal in Fig. 2(a) from its phase. The magnitude is initially set to unity for all ω. The reconstructed signal after 200 iterations is shown in Fig. 2(b). The mean squared error (MSE) between the original and reconstructed signals[1] is non-increasing with each iteration (Fig. 2(c)).

2.2. Reconstruction from magnitude

Unlike phase, the magnitude can not uniquely specify a 1-d sequence. The reason is as follows. If we express the system function as:

$$S(z) = G\prod_{k=1}^{M}(1 - b_k z^{-1}), \tag{2}$$

where G is real, then the square of the magnitude function is expressed as [1]:

$$\begin{aligned} P(z) &= S(z)S^*(1/z^*) \\ &= \prod_{k=1}^{M}(1 - b_k z^{-1})(1 - b_k^* z). \end{aligned} \tag{3}$$

The zeros occur in conjugate reciprocal pairs. Thus the zeros of $S(z)$ (and therefore, the phase) can not be determined by the magnitude alone. Consequently, if the known magnitude (instead of phase) 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[2]. In the case of mixed-phase signals, Van Hove et al. [5] have shown that this ambiguity can be resolved by imposing some phase information (see Section 2.3).

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

2.3. Reconstruction from signed-magnitude

The FT phase is defined by:

$$\phi(\omega) = \arctan(S_i(\omega)/S_r(\omega)) \tag{4}$$

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. [5] show that the magnitude, along with this sign information, provides a unique specification of a finite duration causal sequence. The 'signed-magnitude' is defined as,

$$A(\omega : \omega_o) = \begin{cases} |S(\omega)| & \text{if } -\omega_o \le \phi(\omega) < \omega_o + \pi, \\ -|S(\omega)| & \text{otherwise} \end{cases} \tag{5}$$

where ω_o is an arbitrary number within the interval $[-\pi, \omega_o + \pi)$. Thus, $A(\omega : \omega_o)$ contains information about both the magnitude spectrum and the sign of the real and imaginary parts of the FT. 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 $A_x(\omega : \omega_o) = A_y(\omega : \omega_o)$ for all ω then $x(n) = y(n)$ [5].*

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

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

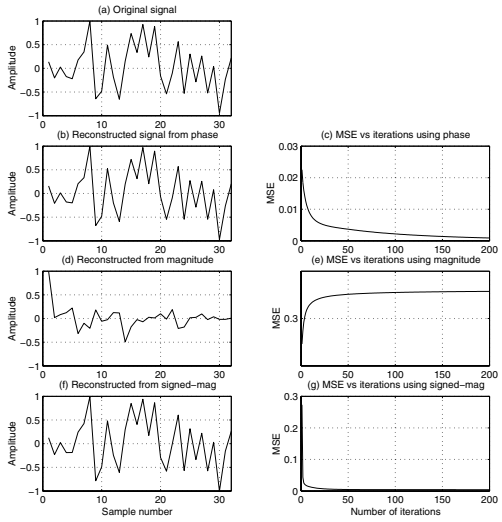

Figure 2: Results of experiments in Section 2. (a) is the original signal. (b), (d) and (f) show reconstructed signals after 200 iterations (these signals are scaled to vary over the same range as the original signal).

3. Reconstruction within the STFT framework

The short-time Fourier transform (STFT) of a speech signal $s(t)$ is given by

$$\begin{aligned} S(\omega, t) &= \int_{-\infty}^{\infty} s(\tau)w(t - \tau)e^{-j\omega\tau}d\tau, \\ &= |S(\omega, t)|e^{j\phi(\omega,t)}, \end{aligned} \tag{6}$$

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

[2] Reconstruction algorithms for minimum phase signals can be found in [2, 3].

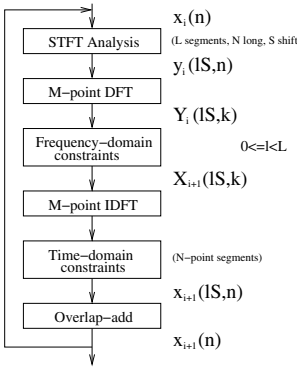

Figure 3: STFT-based iterative reconstruction framework.

where $w(t)$ is a window function of duration T_w, $|S(\omega,t)|$ is the short-time magnitude spectrum and $\phi(\omega,t) = \angle S(\omega,t)$ is the short-time phase spectrum.

Theorems 1 and 2 are also applicable in the context of the STFT [8]. 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. [9]. The framework for the method we use is illustrated in Fig. 3. This method was employed by Griffin and Lim [10] 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 signal is analysed and synthesised in every iteration). In the experiments that follow, we use a rectangular window function of duration T_w=32 ms.

3.1. Reconstruction from STFT phase

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 from each segment. This phase is enforced for every iteration of the reconstruction algorithm. The magnitude is initially set to unity for all frequencies. For all segment shifts, the algorithm converges toward a scaled version of the original signal (Fig. 4(b)&(c)). The convergence is faster for more overlap.

Note that, for the case of reconstruction from short-time phase, there must be at least one sample of overlap between segments. The overlapping sample(s) serve to maintain the energy relationship between adjacent segments. So, even though no energy information is provided to seed the iterative algorithm, the energy contour (albeit scaled) is preserved.

3.2. Reconstruction from STFT magnitude

Griffin and Lim first used the STFT reconstruction framework of Fig. 3 to reconstruct time-scaled versions of a signal from short-time magnitude [10]. Here, we analyse the algorithm for no time-scaling, imposing the known short-time magnitude in each iteration. Short-time phase for each segment is initially randomised. The algorithm does not converge toward the original speech (Fig. 4(d)&(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.

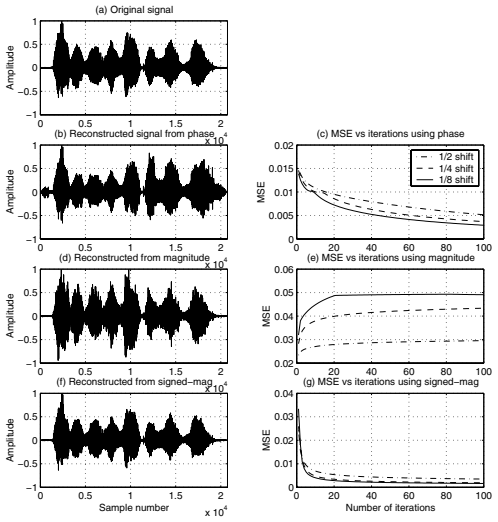

Figure 4: Results of experiments in Section 3. (a) is the original signal used for experiments in Sections 3 and 4. (b), (d) and (f) show reconstructed signals after 100 iterations (these signals are scaled to vary over the same range as the original signal).

3.3. Reconstruction from STFT signed-magnitude

Here, we enforce the known short-time magnitude in addition to the phase sign information (see Section 2.3). Once again, we observe the signal converging, with the rate of convergence increasing, and the error reducing, with more overlap (Fig. 4(f)&(g)).

4. Reconstruction from partial STFT phase

In light of results from previous sections, we note that knowledge of phase is enough for unique signal reconstruction (to within a scale factor), while the same is not true for magnitude. The magnitude must be accompanied by phase-sign information for unique reconstruction. In this section, we explore the use of partial phase information, in the absence of all magnitude information, for intelligible signal reconstruction.

4.1. Reconstruction from STFT phase sign

A similar experiment to that in Section 3.3 is performed. Rather than enforcing sign and magnitude, we only enforce the sign constraint. We observe that MSE does not increase in each iteration (Fig. 5(a)&(b)). More overlap leads to faster reduction in MSE and less error. Informal listening tests indicate that more overlap also leads to better intelligibility. It is interesting that only a small amount of phase information provides for an intelligible signal (although the reconstructed signal is noisy). The increased overlap accommodates, to some extent, for the sparse phase information.

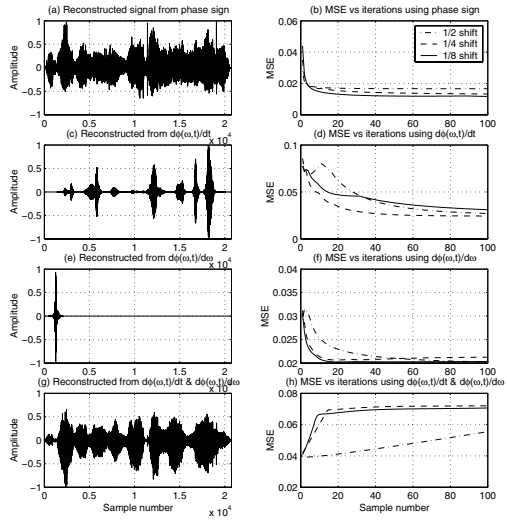

Figure 5: Results of experiments in Section 4. (a), (c), (e) and (g) show the reconstructed signals after 100 iterations (these signals are scaled to vary over the same range as the original signal).

4.2. Reconstruction from time and frequency derivatives of STFT phase

We take the phase of each short-time section and randomise it across frequency, such that $d\phi(\omega,t)/dt$ is preserved. The resulting phase is used in place of the original phase in the reconstruction algorithm. The algorithm does not converge toward the original signal, nor does it provide an intelligible signal (Fig. 5(c)).

In a similar vein, we take the original phase and randomise it across time, preserving the group delay $d\phi(\omega,t)/d\omega$. Again, the reconstruction algorithm does not converge to an intelligible solution (Fig. 5(e)).

Therefore, reconstruction of intelligible speech is not possible from either $d\phi(\omega,t)/dt$ or $d\phi(\omega,t)/d\omega$. This holds true, regardless of the amount of overlap. Figures 5(d) and 5(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 the knowledge of both $d\phi(\omega,t)/dt$ and $d\phi(\omega,t)/d\omega$. In order to do this, we must first reconstruct the phase. Note that the first-segment phase can only be reconstructed to within an additive constant of the original first-segment phase, since all we know about it is $d\phi(\omega,t)/d\omega$. The remaining segments are reconstructed in relation to this segment. Consequently, we cannot recover the original phase. As expected, when attempting to reconstruct the original speech with this phase, the algorithm does not converge (Fig. 5(h)). Regardless of this, a solution that sounds almost exactly like the original speech is provided (Fig. 5(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). Therefore, in the context of the STFT reconstruction framework, when both $d\phi(\omega,t)/dt$ and $d\phi(\omega,t)/d\omega$ are preserved, adequate information is available for intelligible signal reconstruction.

5. Conclusion

We reviewed several signal reconstruction algorithms. Under mild conditions, a finite duration signal can be reconstructed to within a scale factor by its phase (where the phase is determined over the duration of the signal or on a short-time basis). This is not true for magnitude. However, if magnitude is accompanied by some phase information (in the form on Equation 5), then unique reconstruction is possible. This motivated us to explore the use of partial short-time phase information for intelligible signal reconstruction. In the absence of all magnitude information, we observed that (i) intelligible signal reconstruction (albeit noisy) is possible from knowledge of only the phase sign information, and (ii) when both time and frequency derivatives of phase are known, adequate information is available for intelligible signal reconstruction. In the absence of either derivative, an unintelligible signal results.

6. References

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

[2] T.F. Quatieri and A.V. Oppenheim, "Iterative techniques for minimum phase signal reconstruction from phase or magnitude", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-29(6), pp. 1187–1193, Dec. 1981.

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

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

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

[6] L.D. Alsteris and K.K. Paliwal, "Importance of window shape for phase-only reconstruction of speech", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp I-573–I-576, May 2004.

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

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

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

[10] D.W. Griffin and J.S. Lim, "Signal estimation from modified short-time Fourier transform", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-32(2), pp. 236–243, Apr. 1984.