# PERCEPTION OF STOP CONSONANTS IN VCV UTTERANCES RECONSTRUCTED FROM PARTIAL FOURIER TRANSFORM INFORMATION

Jialong He, Li Liu and Günther Palm
Department of Neural Information, University of Ulm, Germany

ABSTRACT - Identification experiments were performed to investigate perception to intervocalic stop consonants in Vowel-Consonant-Vowel (VCV) utterances. The VCV utterances were reconstructed from the following partial Fourier transform information: (1) Long-term Fourier phase spectra; (2) Signed magnitude spectra - Fourier magnitude spectra combined with 1-bit phase spectra. It was shown that the percent of correct identification to the intervocalic consonants was improved from 68% to more than 93% for the first type of stimuli after applying an iterative reconstruction algorithm with enough phase samples. Near-perfect performance could be reached for stimuli reconstructed from signed magnitude spectra. The effects of different initial guesses to the unknown parts and vowel context were discussed.

## I. INTRODUCTION

In the polar representation of a signal's Fourier transform, there are two components: phase and magnitude. For an arbitrary signal, in general, it is not possible to restore the signal only from its Fourier phase or magnitude alone. However, if a signal satisfies certain conditions, it can be completely restored from only its Fourier phase or magnitude. For example, if a signal satisfies the minimum phase condition, i.e., the log magnitude and the phase are related through the Hilbert transform (Oppenheim and Schafer, 1975), it can be exactly recovered from either the magnitude or, to within a scale factor, the phase of its Fourier transform. In many practical applications only partial frequency domain information is available, it is therefore desirable to reconstruct the original signal from this partial information. Signal reconstruction from partial Fourier transform information has drawn great attention to different authors, especially in image processing and other multi-dimensional signal processing areas. Some significant results have been developed. Hayes *et al.* (1980) have proposed a set of conditions under which a finite extent one-dimensional (1-D) or multi-dimensional (MD) sequence is uniquely specified to within a scale factor by the tangent of its Fourier transform phase. Conditions have also been developed under which the signal can be specified to within a translation, sign, and a central symmetry by Fourier magnitude (Hayes, 1982). Hove *et al.* (1983) have extended these conclusions and given another set of restrictions under which both 1-D and MD are completely specified by their signed magnitude (magnitude with one bit phase information at each frequency component). Curtis *et al.* (1985) have developed some theoretical results which state conditions under which two-dimensional signals are uniquely specified to within a scale factor by the sign of the real part of the Fourier transform. Unlike that in image research, little attention has been focused on the reconstruction of a speech signal from its partial Fourier information. One reason is that the phase spectrum was generally accepted as less or unimportant for an acoustic signal. However, Oppenheim and Lim (1981) studied the role of Fourier phase in the representation of a speech signal and compared it with that of Fourier magnitude. They demonstrated that, in many contexts, the phase contains very important information. Recent studies also shown that the phase is very important for perceiving consonants when analysis window size is relatively larger (Liu *et al.*, 1992). There are also a number of practical applications, e.g., speech coding, in which phase may play an important role. In this paper we present some experimental results of identifications to the intervocalic stop consonants in Vowel-Consonant-Vowel (VCV) signals reconstructed from phase spectra or signed magnitude spectra. An accelerated iterative algorithm was also discussed. It was shown that the consonant identification rate to the phase-only stimuli improved from 68% to more than 93% after applying the iterative algorithm. Near-perfect identification performance could be reached for the stimuli reconstructed from signed Fourier magnitude spectra.

## II. ALGORITHM

Despite the fact that, under certain conditions, a signal can be uniquely specified by its partial Fourier transform information, practical and efficient algorithms to reconstruct it from this partial information generally demand additional constraints. One of the magnitude-retrieval algorithms called closed-form

solution was proposed by Hayes (1980). Since this method needs to solve a set of N-1 linear equations, where N is the total number of unknown speech samples, it will lead to numerical problems and suffer severe roundoff errors when N becomes very large. The most commonly used method for signal reconstruction from partial information is an iterative technique developed by Gerchberg and Saxton (1972) and Fienup (1978) for reconstructing a signal from magnitude information and by Quatieri (1981) for reconstructing a signal from its phase under the assumption that it is a minimum phase signal. In the iterative solution the transformations between time and frequency domains are repeatedly invoked. In each domain the known information or the constraints about the signal are incorporated into the current estimation. The basic algorithm is shown schematically in figure 1.

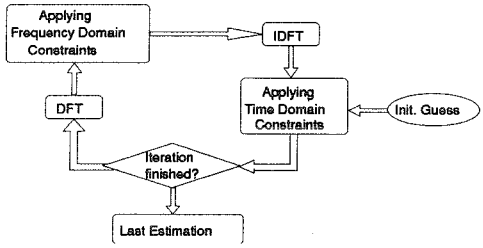


Figure 1. Block diagram of the basic iterative algorithm for reconstruction a signal from partially known information.

To speed up the rate of convergence, we employed a modified version of the basic algorithm with the adaptive relaxation coefficient (Hayes, 1982). The following is detail of the algorithm. For the sake of simplicity, we first define four operators:

(1) $\mathcal{F}$ is a discrete Fourier transform operator: $\mathcal{F}\{S(n)\} = \{A(k)\ ,\ P(k)\}$
(2) $\mathcal{F}^{-1}$ is an inverse discrete Fourier transform operator: $\mathcal{F}^{-1}\{A(k)\ ,\ P(k)\} = \{S(n)\}$
(3) $\mathcal{T}$ is a time domain constraint operator: $\mathcal{T}\{S(n)\} = \{S_t(n)\}$
(4) $\mathcal{Q}$ is a frequency domain constraint operator: $\mathcal{Q}\{A(k)\ ,\ P(k)\} = \{A_q(k)\ ,\ P_q(k)\}$

where $S(n)$ is a speech signal, $A(k)$ and $P(k)$ are its magnitude and phase spectra. $S_t(n)$ is the signal that satisfies the time domain constraint $\mathcal{T}$. $A_q(k)$ and $P_q(k)$ are the magnitude and phase spectra that satisfy the frequency domain constraint $\mathcal{Q}$.

**STEP 1: Initial guesses to unknown parts**
To reconstruct a signal $S(n)$ from its phase spectrum, we should make an initial guess to the unknown magnitude spectrum $A^0(k)$. Similarly, to reconstruct a signal from its signed magnitude spectrum, we select an initial guess to the unknown phase spectrum $P^0(k)$. The initial guess $S^0(n)$ can be obtained by applying the inverse discrete Fourier transform $\mathcal{F}^{-1}$ to the magnitude and phase spectra.

**STEP 2: Applying time and frequency domain constraints**
Suppose that $S^k(n)$ is the estimation for $S(n)$ at the *k*th iteration and the intermediate result of *(k+1)*th estimation is defined by:

$$\tilde{S}^{k+1}(n) = \mathcal{F}^{-1}\mathcal{Q}\mathcal{F}\mathcal{T}\{S^k(n)\} \qquad (1)$$

where the four operators are defined as above and applied to $S^k(n)$ in precedence from right to left. In fact, this is the basic iteration procedure shown in figure 1.

**STEP 3: Calculating adaptive relaxed coefficient $\alpha^k$**
An adaptive relaxation technique for the iterative algorithm is defined as:

$$S^{k+1}(n) = (1-\alpha^k)S^k(n) + \alpha^k \tilde{S}^{k+1}(n) \quad (2)$$

where $\alpha^k$ is known as the relaxed parameter. One possibility for increasing the rate of convergence is to find a value of $\alpha^k$ which minimize the error between the *(k+1)*th estimation and the time domain constraints. Suppose that $S(n)$ is a N point signal and the first M (<N) points are known, e.g., time limited. The mean-square error is defined by:

$$E = \sum_{n=0}^{M} \left(S^{k+1}(n) - S(n)\right)^2 \quad (3)$$

Substituting $S^{k+1}(n)$ with equation (2) and setting $\frac{dE}{d\alpha^k} = 0$:

$$\tilde{\alpha}^k = \sum_{n=0}^{M} \left(S(n) - S^k(n)\right)\left(\tilde{S}^{k+1}(n) - S^k(n)\right) / \sum_{n=0}^{M} \left(\tilde{S}^{k+1}(n) - S^k(n)\right)^2 \quad (4)$$

Since the value $\tilde{\alpha}^k$ obtained from the above expression does not guarantee the iterative algorithm to be nonexpansive, we selected the relaxed parameter as:

$$\alpha^k = \begin{cases} 0 & \tilde{\alpha}^k < 0 \\ \tilde{\alpha}^k & 0 \le \tilde{\alpha}^k \le 2 \\ 2 & \tilde{\alpha}^k > 2 \end{cases} \quad (5)$$

The estimation for $S(n)$ at *(k+1)*th iteration can be calculated from equation (2).

**STEP 4: If *(k+1)* < desired iteration number; go to STEP 2**

## III. EXPERIMENT

### A. Speech signals

The original signals consisted of all 12 combinations of 6 stop consonants /b, p, t, d, k, g/ with two vowels /a, i/ and were recorded in a sound-isolated booth. The two vowels in each VCV syllable were identical (e.g. /apa/, /iki/). The VCVs were read by 5 untrained German speakers in isolation with approximately equal stress on the two syllables and were digitized at 16 kHz sampling rate with 16-bit precision. The length of an original VCV signal was about 512 ms (8192 samples).

### B. Subjects

Twelve paid students served as subjects. All were native German speakers with no known history of either speech or hearing disorders. They were tested in small groups (2-4 subjects) in a sound-isolated booth. All stimuli were presented binaural over headphones at a comfortable listening level. The subjects were instructed to write down the perceived intervocalic stop consonant in the designated blank on the response sheet.

### C. Initial guesses

To test the signals reconstructed from their phase spectrum, three kinds of initial guesses to the unknown magnitude $A^0(k)$ were used in our experiments:

(1) $A_1^0(k)$ : a constant spectrum;
(2) $A_2^0(k)$ : an averaged spectrum from uncorrelated utterances;
(3) $A_3^0(k)$ : an averaged spectrum from /aCa/ utterances;

Another type of stimuli were reconstructed from the signed magnitude spectra. The initial guess to the magnitude and phase spectra were:

$$A^0(k) = |A_s(k)|$$

$$P^0(k) = \begin{cases} 0 & A_s(k) < 0 \\ \pi & A_s(k) \geq 0 \end{cases} \quad (6)$$

where $A_s(k)$ was the signed magnitude spectrum.

**D. Time domain constraints**

We tested two different time domain constraints $\mathcal{T}1$ and $\mathcal{T}2$. (1) $\mathcal{T}1$: signals were time limited within 512 ms, i.e. values outside of this range were zero. Since a time signal is periodically repeated by discrete Fourier transform, to prevent time alias, we have to use the DFT with more than 8192 points (in our case 16384 points). In another word, we should know more phase or signed magnitude samples. When $\mathcal{T}1$ was applied to a signal, the values outside the range of 8192 points would be set to zero. (2) $\mathcal{T}2$: We supposed that the first 4096 samples of a VCV signal, normally the first vowel, were known. In this case, we could use 8192 point DFT. When applying $\mathcal{T}2$ to a signal, its first 4096 points would be replaced with the known samples of the VCV.

**E. Frequency domain constraints**

Two kinds of frequency domain constraints were used in our experiments. (1) $\mathcal{Q}1$: signals' phase spectra were known. When applying $\mathcal{Q}1$ to a spectrum, the phase would be replaced with the known phase spectrum and the magnitude spectrum kept unchanged. (2) $\mathcal{Q}2$: signals' signed magnitude spectra were known. When $\mathcal{Q}2$ was applied to a spectrum, the magnitude and phase spectra were obtained from:

$$A_q(k) = |A_k(k)|$$

$$P_q(k) = \begin{cases} P^k(k) & (A_s(k) < 0,\ -\pi/2 \leq P^k(k) < \pi/2) \\ & \text{or } (A_s(k) \geq 0,\ \pi/2 \leq P^k(k) < 3\pi/2) \\ P^k(k) + \pi & \text{other} \end{cases} \quad (7)$$

where $A_s(k)$ is the known signed magnitude spectrum, $P^k(k)$ is the phase spectrum before applying $\mathcal{Q}2$. For clarity, Table 1 summarized our 10 categories of stimuli according to the different initial guesses to the unknown parts and constraints. Stimuli in category J were original signals and were included here for comparisons.

## IV. RESULT

Figure 2 shows the percent of correct identification to intervocalic consonants as a function of categories under two different vowel contexts (/a/ and /i/). The identification rate for stimuli synthesized by combining phase spectra with a flat magnitude spectrum, *without applying iteration algorithm*, was also included in Figure 2 (denoted as 'Flat') for comparisons. It was excerpted from another experiment (Liu et al. 1992). By examining the stimuli reconstructed from phase (categories A-F), it is easy to see that the identification rates for the stimuli with time domain constraint $\mathcal{T}2$ are relatively higher than that with $\mathcal{T}1$. An explanation is that when reconstructing a signal with $\mathcal{T}1$ (as in category A, C and E) we supposed the samples outside the range (0, 8191) were zero, on the other hand, by using $\mathcal{T}2$ we knew a part of nonzero samples which contribute to both magnitude and phase spectra, consequently, this would be more helpful to restore the original magnitude spectrum than in the case of $\mathcal{T}1$. Besides, data from acoustic measurements and from perception experiments indicate that both formant transitions from the first vowel to the closure (VC transitions) and transitions out of the closure into the second vowel (CV transitions) contribute to identifications of the intervocalic stop. If the first half of a VCV speech signal is known, as in the case of category B, D and F, a part of original information about the consonant is actually preserved. The intervocalic stop consonants may

be identified with a relatively higher performance even if the retrieval process was not very successful. In figure 2, it has also been shown that the iterative procedure improved the identifiability of consonants significantly for phase-only stimuli than that without iteration (labeled as Flat).

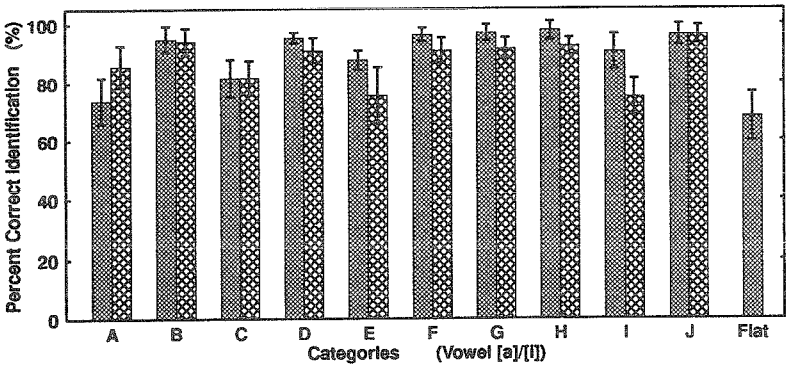


Figure 2. The percentages of correct consonant identification and their standard deviations as a function of the categories.

The vowel context also has some effects on the consonant identification when the initial guess to the unknown magnitude spectrum is flat (category A) or averaged from all /aCa/s (category E). The effect of vowel context in the case of category E is easy to understand since the initial guess to magnitude spectrum in category E contained more information about vowel /a/ than /i/, so that /aCa/ signals should be restored better than /iCi/. However, no convincible explanation was found to the fact that vowel context had an effect in the case of category A in which the initial guess to the magnitude was a flat spectrum. There was no difference on identification performances to the original signals under different vowel contexts (category J). The initial guess to the unknown magnitude was found to be another factor which influences the identification rates. The /aCa/s could be identified much better when the initial magnitude guess contained information about vowel /a/ (category E) than when it didn't (category A and C). Correspondingly, the lowest identification rate was obtained for the /iCi/s when the initial magnitude guess contained information contradicting the vowel context /i/. Near perfect identification performance was obtained for the stimuli synthesized from signed magnitude spectra (category G and H). The identification rates was significantly improved after applying the iterative procedure in comparison with the stimuli constructed by simply combining the known magnitude spectra with 1-bit phase spectra (category I).

## V. REFERENCE


Curtis, S. R., Oppenheim, A. V. and Lim, J. S. (1985) "Signal reconstruction from fourier transform sign information," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-33, pp. 643-657.

Fienup, J. R. (1978) "Reconstruction of an image from the modulus of its Fourier transform," Opt. Lett., Vol. 3, pp. 27-29.

Gerchberg, R. W. and Saxton, W. O. (1972) "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik, Vol. 35, pp. 237-246.

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

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, vol. ASSP-30, pp. 140-154.

van Hove, P.L., Hayes, M.H., Lim, J.S. and Oppenheim, A.V. (1983). "Signal reconstruction from signed fourier transform magnitude," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-31, pp. 1286-1293.

Liu, L., He, J. L. and Smit, A. C. (1992) "The importance of phase in the perception of intervocalic stop consonants," J. Acoust. Soc. Am. 92, pp. 2340.

Oppenheim, A.V. and Schafer, R.W. (1975). Digital Signal Processing. Englewood Cliffs, NJ: Presentice-Hall.

Oppenheim, A. V. and Lim, J. S. (1981) "The importance of phase in signals," Proc. IEEE, vol. 69, pp. 529-541.

Oppenheim, A. V., Hayes, M. H. and Lim, J. S. (1982) "Iterative procedures for signal reconstruction from fourier transform phase," Optical Engineering, vol. 21, pp. 122-127.

Oppenheim, A. V., Lim, J. S. and Curtis, S. R. (1983) "Signal synthesis and reconstruction from partial fourier domain information," J. Opt. Soc. Am. , Vol. 73, pp. 1413-1420.

Quatieri Jr., T. F. and Oppenheim, A. V. (1981) "Iterative Techniques for minimum phase signal reconstruction from phase or magnitude," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp. 1187-1193.

| Category | Initial guess | time constraint | freq. constraint | FFT/IFFT size | Known information |
|---|---|---|---|---|---|
| A | $A_1^0(k)$ | $\mathcal{T}1$ | $\mathcal{Q}1$ | 16384 | 8192 point phase |
| B | $A_1^0(k)$ | $\mathcal{T}2$ | $\mathcal{Q}1$ | 8192 | 4096 phase, 4096 signal |
| C | $A_2^0(k)$ | $\mathcal{T}1$ | $\mathcal{Q}1$ | 16384 | 8192 point phase |
| D | $A_2^0(k)$ | $\mathcal{T}2$ | $\mathcal{Q}1$ | 8192 | 4096 phase, 4096 signal |
| E | $A_3^0(k)$ | $\mathcal{T}1$ | $\mathcal{Q}1$ | 16384 | 8192 point phase |
| F | $A_3^0(k)$ | $\mathcal{T}2$ | $\mathcal{Q}1$ | 8192 | 4096 phase, 4096 signal |
| G | $P^0(k)$ | $\mathcal{T}1$ | $\mathcal{Q}2$ | 16384 | 8192 point signed magnitude |
| H | $P^0(k)$ | $\mathcal{T}2$ | $\mathcal{Q}2$ | 8192 | 4096 signed magnitude, 4096 signal |
| I | $P^0(k)$ | none | none | 8192 | 4096 signed magnitude |
| J | none | none | none | 8192 | 8192 sample |

Table 1. Summary of 10 categories of stimuli