On the Importance of Phase in Human Speech Recognition

Guangji Shi, *Student Member, IEEE*, Maryam Modir Shanechi, and Parham Aarabi, *Member, IEEE*

***Abstract*—In this paper, we analyze the effects of uncertainty in the phase of speech signals on the word recognition error rate of human listeners. The motivating goal is to get a quantitative measure on the importance of phase in automatic speech recognition by studying the effects of phase uncertainty on human perception. Listening tests were conducted for 18 listeners under different phase uncertainty and signal-to-noise ratio (SNR) conditions. These results indicate that a small amount of phase error or uncertainty does not affect the recognition rate, but a large amount of phase uncertainty significantly affects the recognition rate. The degree of the importance of phase also seems to be an SNR-dependent one, such that at lower SNRs the effects of phase uncertainty are more pronounced than at higher SNRs. For example, at an SNR of -10 dB, having random phases at all frequencies results in a word error rate (WER) of 63% compared to 24% if the phase was unaltered. In comparison, at 0 dB, random phase results in a 25% WER as compared to 11% for the unaltered phase case. Listening tests were also conducted for the case of reconstructed phase based on the least square error estimation approach [11]. The results indicate that the recognition rate for the reconstructed phase case is very close to that of the perfect phase case (a WER difference of 4% on average).**

***Index Terms*—Phase analysis, phase effect, phase reconstruction, speech recognition.**

I. Introduction

CURRENT state-of-the-art speech recognition system can achieve high recognition accuracy rates (>90%) in noise free environments [1]. However, their performance significantly degrades in adverse situations when noise and/or reverberation are present [2]. Since most environments do contain noise and reverberation, a solution must be found to enable robust and accurate speech recognition in all practical situations. This solution can fall under any of the following two categories.

The first method to solve this problem is to enhance speech by removing noise prior to recognition. There have been numerous algorithms [1], [3], commonly employing multiple microphones, which have reported significant gains in recognition accuracy rate.

Another strategy in improving speech recognition systems is to use a more complicated recognition model; one which takes in and processes more information. Most state-of-the-art speech recognition systems only utilize the magnitude of the Fourier transform of the time-domain speech segments [2]. This means that the corresponding Fourier transform phases are discarded. Several studies have indicated that it may be a fruitful effort to directly model and incorporate the phase into the recognition process [4], [5], [12], [13]. Furthermore, several studies have shown the perceptual importance of phase in speech coding [7]–[9].

In this paper, we will attempt to answer a simple question: how much does phase really matter in the recognition process? We will answer this question experimentally by performing human speech recognition experiments with different amounts of uncertainty in the phase and with different SNR conditions.

II. Problem Statement and Prior Work

In general, the signal received by a human ear or by a microphone can be expressed as [6]

$$x(t) = h(t) * s(t) + n(t) \tag{1}$$

where $s(t)$ is the speech signal of interest, $h(t)$ is the impulse response associated with the source and receiver, and $n(t)$ is the noise signal. In the frequency domain, this is represented as

$$X(\omega) = H(\omega)S(\omega) + N(\omega) \tag{2}$$

where the capital letters are all Fourier transforms of their lowercased time domain representations.

In practice, we can only obtain a sampled finite-duration segment of $x(t)$, which leads to a discrete frequency representation [through the discrete Fourier transform (DFT)]. In other words, the function $X(\omega)$ is only known at a discrete set of values from $\omega = -\pi F_s$ to $\omega = \pi F_s$ in $2\pi F_s/N$ steps. Clearly, $X(\omega)$ is a complex number with a magnitude and a phase (i.e., $X(\omega) = |X(\omega)|e^{j\angle X(\omega)}$). With a few exceptions, current state-of-the-art speech recognition systems only utilize $|X(\omega)|$ and ignore the phase [2]. The phase $\angle X(\omega)$, however, does have potential uses and as a result has gained greater focus from the research community [3]–[5]. In [3], for example, the difference in phase for a pair of microphones was used to attenuate noise, yielding a substantially higher (about 20%) speech recognition accuracy rate. Two questions clearly arise at this point: how much does phase matter in the recognition of speech, and what is the optimal method for incorporating phase into the speech recognition process? In this paper, we will focus on the former question.

Much of the work on human perception of phase is done in speech coding [7], [9]. One primary requirement in speech coding is to use the minimal amount of data to represent the original signal while keeping the coding distortions below the

Manuscript received November 16, 2004; revised July 29, 2005. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Arun C. Surendran.

The authors are with the Department of Electrical and Computer Engineering, University of Tornoto, Toronto, ON M5S 3G4, Canada (e-mail: guangji@comm.utoronto.ca; shanechi@mit.edu; parham@ecf.utoronto.ca).

Digital Object Identifier 10.1109/TSA.2005.858512

threshold of human perception. Traditionally, only the magnitude information is coded. An example of this is sinusoidal coding, where phase is estimated at the decoder based on a minimum-phase assumption [7]. However, losing the phase information usually results in the degradation of the reconstructed speech (i.e., it sounds less natural) [9].

In the past, a number of studies were conducted on the importance of phase in human perception [8]–[10], [13]. In [9], the characteristics of human phase perception were analyzed in terms of just noticeable difference (JND) of phase. The results indicated that human perception of phase varies with frequency, especially for low pitched speakers. This dependence is particularly strong in the midfrequency range (1–3 KHz). In [10], the role of phase on the human perception of intervocalic stop consonants was investigated. It was shown that the phase spectra play an important role in specifying a stop consonant. In [13], experiments were conducted to investigate the usefulness of phase spectrum in human speech recognition. The results showed that phase spectrum can significantly contribute to speech intelligibility.

Some interesting results have also been obtained by using phase features for speech recognition. In [4], time-domain based phase features were extracted and employed as discriminative features. In [12], frequency related features extracted from the phase of speech were used for recognition of vowels. In [5], the short term Fourier phase-spectrum was used for phase feature extraction. The extracted features were then used together with the Mel frequency cepstral coefficients (MFCC) for improved digit recognition.

In this paper, we will consider the most successful speech recognition system that currently exists, namely, the human speech recognition system. We shall then observe the relative speech recognition rates for humans when the phase of the speech signal is altered, compared to the unaltered phase case.

III. Speech Recognition Experiment

To evaluate the role of phase in the recognition of speech by humans, an experiment was conducted with 18 listeners and six speakers (three male speakers and three female speakers). The audio signals for the speakers were obtained from the Audio-Visual Data Corpus from Carnegie Mellon University. This database consists of 78 preselected words recorded in a soundproof studio environment with a sampling frequency of 44.1 Khz for each speaker.

For the purposes of the phase importance experiment, the recorded sound files were segmented such that each segment contained a single word. Gaussian noise, at levels of 20 dB, 0 dB, -5 dB, or -10 dB was added to each segment. At each SNR, the phase of the original speech signal was modified for all frequencies using the following formula:

$$\angle\hat{X}(\omega) = (1-\alpha)\angle X(\omega) + \alpha\phi \tag{3}$$

where ϕ is a randomly generated phase, uniformly distributed between $-\pi$ to π, and α is the phase noise factor ranging from 0 (perfect phase) to 1 (completely random phase). As a result, the altered signal would have the following relation to the original signal:

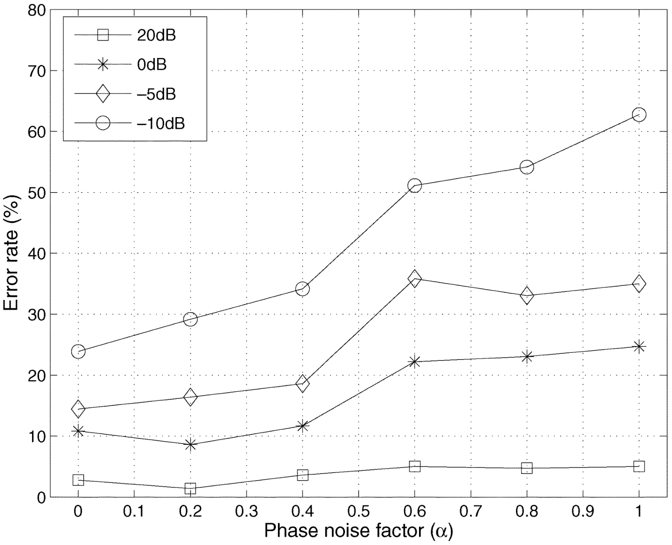


Fig. 1. Word recognition error rate versus phase noise factor at different SNRs.

$$\hat{X}(\omega) = |X(\omega)|e^{j\left((1-\alpha)\angle X(\omega)+\alpha\phi\right)}. \tag{4}$$

This modification was performed on the DFT of half overlapped Hanning-windowed 512-sample segments. This was done to reflect the processing that is performed in current speech recognition systems. After the modifications, $\hat{X}(\omega)$ for each segment was transformed (using the Inverse DFT) back into the time-domain, half overlapped, and added to form the phase-disrupted signal.

The listening experiments were carried out at each value of α and at each SNR for each of the 18 listeners. Each listening trial consisted of 20 randomly selected words. Before the actual test, the volume of the speaker was adjusted to a comfortable level for each listener. Hence, with 20 words per trial, three different SNR conditions, and six different values of α (i.e., $\alpha = 0, 0.2, \ldots, 1.0$), a total of 360 words were tested for recognition by each listener. Half of the listeners were presented with the order of 0, -5, and -10 dB. The other half of the listeners were presented with the exactly opposite order. For the phase noise factor, all listeners were first presented with the $\alpha = 0$ (no phase noise) case, and then α was increased gradually to 1 with a step size of 0.2.

Fig. 1 shows the average word recognition error rates over all listeners with different phase noise factors and with four different SNRs (-10, -5, 0, and 20 dB). As shown in this figure, phase noise can change the recognition rate by as much as 39% when the input SNR is -10 dB. This difference becomes smaller (about 15%) when the input SNR is 0 dB. At 20 dB, the influence of phase on speech recognition rate is small despite the fact that one can still perceive the existence of phase noise. Clearly, the question of the importance of phase seems to be an SNR-dependent one.

Since the effect of phase on recognition accuracy is small at high SNRs (as shown by the 20 dB case), we will focus on the low SNR regions. Fig. 1 also shows that the recognition error rate curves for the three low SNR values (-10, -5, and 0 dB) are

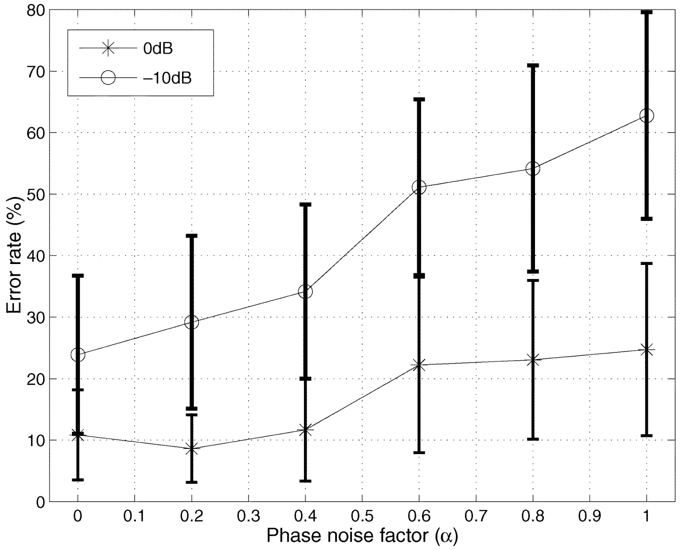


Fig. 2. Error bar plot of word error rate for the -10 dB and 0 dB cases.

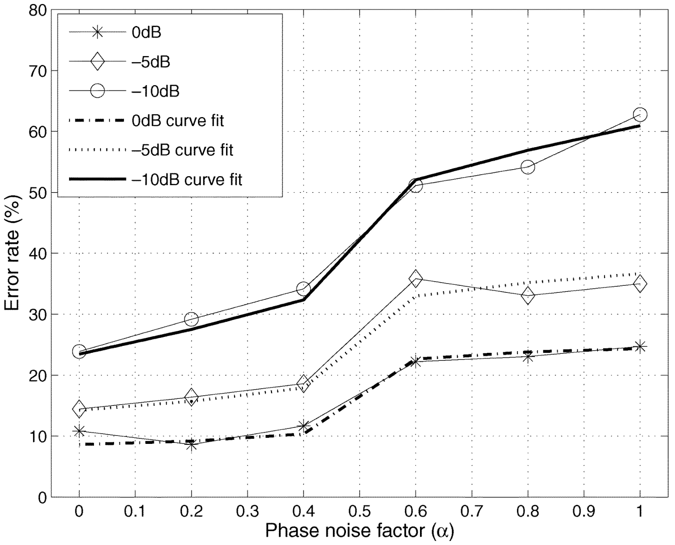


Fig. 3. Approximation of the effect of uncertainty in phase on the recognition error rate.

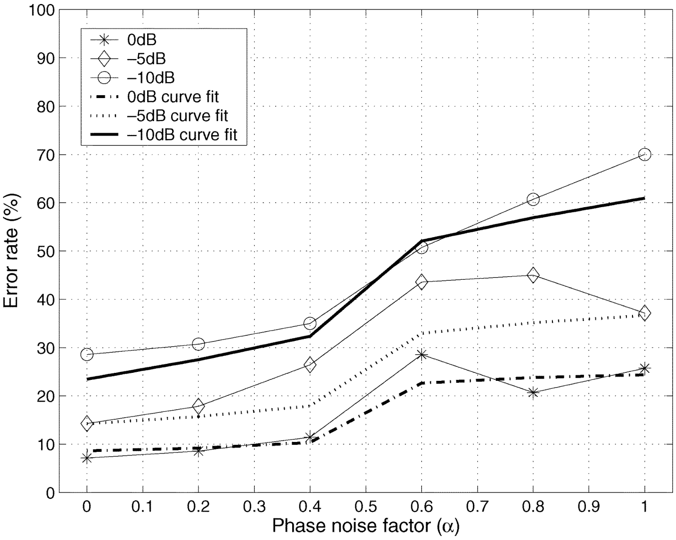


Fig. 4. Average recognition errors in the case of male speakers.

fairly consistent. The results for these three cases indicate that small phase noise factors ($\alpha < 0.4$) do not significantly affect the recognition rate. However, for phase noise factors greater than 0.6, there is a consistent and significant increase of error rate. Furthermore, there is a jump in error rate when the phase noise factor changes from 0.4 to 0.6.

Fig. 2 shows the error bar plot for the -10 and 0 dB cases. The height of each bar represents one standard deviation above and one standard deviation below the mean. It is clear that there is a variation in the perception of phase among different listeners. This is representative of the different listening capability, command of English, focus of the different listeners, and etc. To minimize the effects due to factors other than phase noise, we will only consider the averaged results over multiple speakers. Despite of the variation among different listeners, the general trend which was observed in the means remains the same: a phase noise factor beyond 0.5 significantly affects the recognition process, especially at lower SNRs.

IV. Modeling the Effect of Phase

In the previous section, we have observed that the effect of phase is SNR dependent. Furthermore, the phase effect is fairly consistent at low SNRs: There is a sigmoidal jump of error rate between $\alpha = 0.4$ and $\alpha = 0.6$. Otherwise, the error rate curves are approximately linear. To facilitate further analysis on the effect of phase, we propose to use the expression below to model the experimental results we obtained at low SNRs (shown in Fig. 1).

$$\psi(R,\alpha) = e^{-(R+24.5)/10} + \alpha e^{-(R+18)/5} + \frac{0.13e^{-R/35}}{1+e^{-30(\alpha-0.5)}} \tag{5}$$

where α is the phase noise factor and R is the SNR in dB. This model is obtained through curve fitting of the results obtained at 0, -5, and -10 dB shown in Fig. 1. There are three terms in this equation. The first two terms are used to model the linear relation between the phase noise factor and the recognition error rate. The third term is used to model the sigmoidal jump of the error rate curves. The numerator of the third term controls the amount of the jump. The model in (5) is not meant to be a general model. It is only a rough model that we will use to conduct further analysis. Fig. 3 compares the result of this approximation with the test result. Clearly, at the experimented SNRs (-10, -5, 0 dB), the proposed model matches the results well.

Of the 18 randomly selected experiments, 7 had male speakers and 11 had female speakers. Figs. 4 and 5 show the average recognition errors for six different values of α for the case of male speakers and female speakers, respectively.

The experimental results obtained above are summarized in Fig. 6. This figure shows the recognition error rate as a function of SNR for the perfect phase case and for the completely random phase case. We have also plotted the two data points at 20 dB to provide a more complete picture. This figure, which is a culmination of the results so far, shows the *worst-case* increase

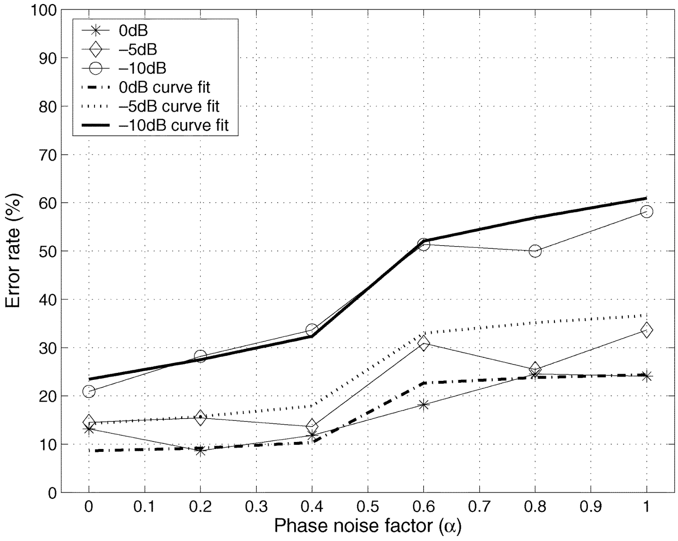


Fig. 5. Average recognition errors in the case of female speakers.

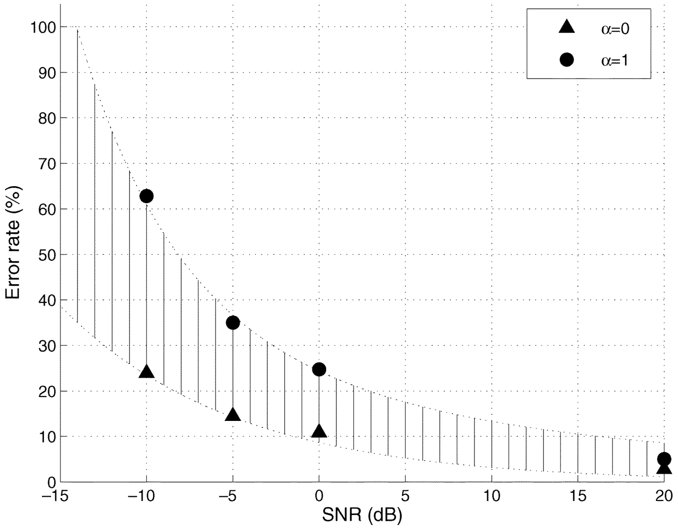


Fig. 6. Average intervals of phase influence on speech recognition at different SNRs: The modeling results are shown as dotted curves, and the 18-listener experimental results are shown as discrete points. We have also plotted the two data points at 20-dB SNR to provide a more complete picture.

in recognition error rate due to uncertainty in phase for the ultimate speech recognition system (i.e., that of humans). Hence, as speech recognition systems improve to better match that of humans, our results define an experimental upper bound on the recognition accuracy rate gain that can be obtained by modeling phase (versus not modeling the phase). Furthermore, we can observe that this gain is SNR dependent: it is much higher at lower SNRs than at higher SNRs. In other words, phase matters much more at lower SNRs than it does at higher SNRs.

Based on the proposed model, a contour plot was made relating speech recognition error rate (ψ) in humans to the phase noise factor and the SNR (Fig. 7). This plot shows more clearly than before that phase has almost a bimodal effect on the recognition rate: it is either not very significant or very significant, depending on whether the phase noise factor is greater than or less than 0.5. The SNR has a clear and direct effect on the recognition rate, but this effect is more gradual which leads to higher recognition error rates at lower SNRs.

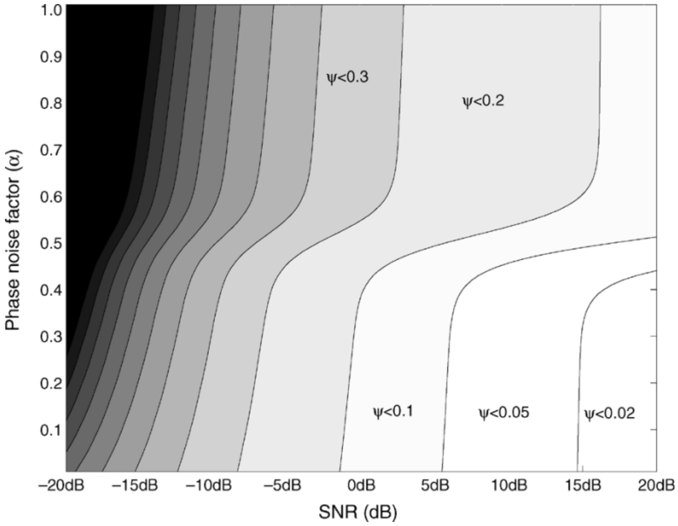


Fig. 7. Contour plot of recognition accuracy as a function of SNR and phase noise factor.

V. Phase Estimation From Spectrogram

Having observed the effect of phase on human speech recognition, we proceed to evaluate the effectiveness of phase restoration techniques such as the one in [11].

In [11], Griffin proposed a least square error estimation (LSEE) approach for signal estimation from the short time Fourier transform (STFT) magnitude. Let $x(n)$ and $\hat{x}(n)$ denote a speech segment and its estimate, then the algorithm in [11] minimizes the following distance function:

$$D = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(|X(kL,\omega)| - |\hat{X}(kL,\omega)| \right)^2 d\omega \quad (6)$$

where $X(kL,\omega)$, $\hat{X}(kL,\omega)$ are the STFTs of $x(n)$, $\hat{x}(n)$ at frequency ω and for time segment k respectively, and L is the sampling period of $X(n,\omega)$ (or $\hat{X}(n,\omega)$) in the variable n. The solution to (6) yields an iterative procedure. Assume $\hat{x}_i(n)$ is the estimate of $x(n)$ after the ith iteration, then the estimate at the next iteration is given by

$$\hat{x}_{i+1}(n) = \frac{\sum\limits_{k=-\infty}^{\infty} h(kL-n) \int_{\omega=-\pi}^{\pi} \hat{X}_{i+1}(kL,\omega) e^{j\omega n} d\omega}{\sum\limits_{k=-\infty}^{\infty} h^2(kL-n)} \quad (7)$$

where $\hat{X}_{i+1}(kL,\omega) = |X(kL,\omega)| \angle \hat{X}_i(kL,\omega)$, $h(n)$ is the window used, and $\hat{X}_i(kL,\omega)$ is the STFT of $\hat{x}_i(n)$.

We used the LSEE algorithm above to preprocess the entire database (one word at a time) at different SNR levels. Gaussian noise was added to each word segment the same way as before. Then, each word segment was partitioned into half overlapping 512-sample segments which were windowed by a Hanning window. Each estimation process took 100 iterations.

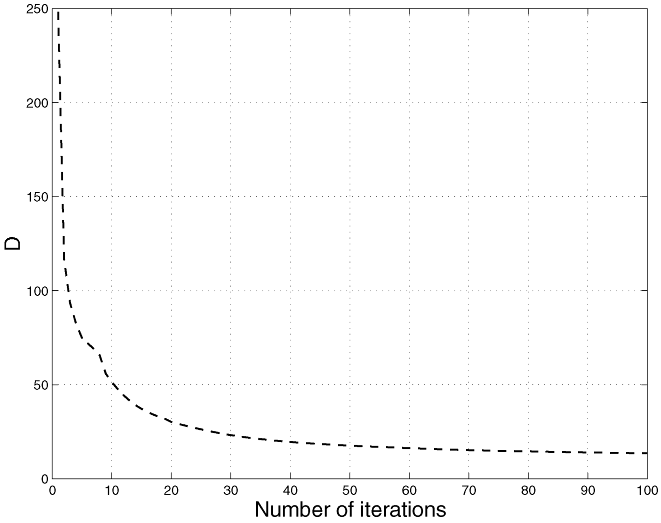


Fig. 8. Distance measure (D) of (6) versus number of iterations using the LSEE phase restoration algorithm.

To demonstrate the convergence behavior of the LSEE algorithm, we have plotted a convergence curve of a randomly selected word segment in Fig. 8. As shown in this figure, the distance measure (D) of (6) decreases rapidly in the first few iterations. After that, D decreases more slowly. By listening to the signals with reconstructed phase, good results can be obtained when the number of iterations exceeds 30. This is supported by the convergence curve in Fig. 8 since D does not change much after iteration 30. We have set the number of iterations to 100 to utilize the full capacity of the LSEE algorithm.

VI. Phase Restoration Effects on Speech Recognition

To investigate the corresponding phase noise factor for the phase restored signals, a second set of experiments was conducted with 18 listeners. The audio signals used were identical to the ones used in the first set of experiments. Again Gaussian noise at levels of 0, -5, and -10 dB was added to each segment. This time for each SNR level, phase noise factors of $\alpha = 0$ (corresponding to perfect phase) and $\alpha = 1$ (corresponding to completely random phase) were studied. By evaluating these two extreme cases, we could test the model in (5). Also at each SNR level, the recognition experiments were repeated for the reconstructed speech signals from the method of [11].

Table I shows the average WERs over all 18 speakers for the cases of perfect phase, reconstructed phase and complete random phase at different values of SNR. Before analyzing the effect of phase reconstruction, we first verify our model in (5) using the new test results obtained from the two extreme cases (perfect phase and completely random phase). The filled circles in Fig. 9 show the corresponding recognition rates for each of the SNR levels and for cases of $\alpha = 0$ and $\alpha = 1$. Also shown in the figure are the estimated curves using (5). We can see that the estimated curves are able to estimate the new data points well except for the case of -10 dB and $\alpha = 0$. This is likely due to anomaly generated from listeners in the second set of experiments. Nevertheless, the test result for this particular case is still within one standard deviation of the estimated value. Overall, the general trend does hold.

TABLE I
WER Comparison of Perfect Phase, Reconstructed Phase, and Random Phase

	Perfect Phase	Reconstructed Phase	Random Phase
0dB	9.44%	13.06%	22.22%
-5dB	16.39%	20.56%	40.00%
-10dB	33.61%	37.50%	64.72%

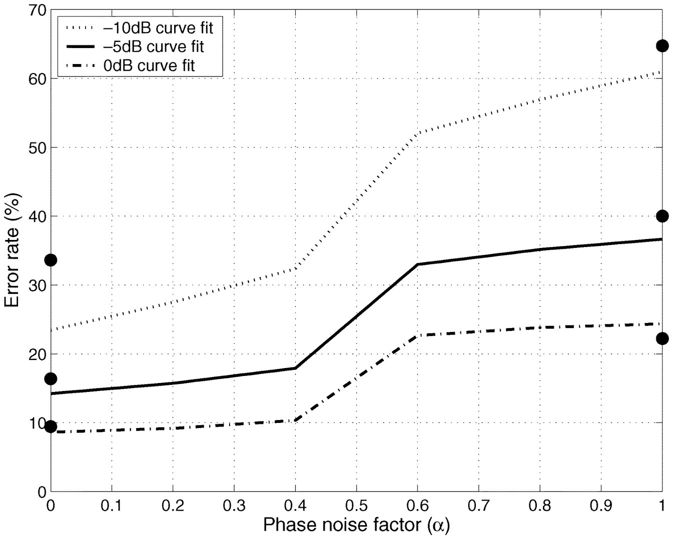


Fig. 9. Approximation of the results obtained in the second set of experiments using the model in (5). The curves show the estimated error rates using (5). The solid dots show the new set of recognition rates found in the second set of experiments for the two extreme cases ($\alpha = 0$ and $\alpha = 1$).

The phase noise factors of the reconstructed speech at the three different SNRs can be estimated using their error rates obtained in the second experiment and the model in (5). Based on the model, word error rates of 37.50%, 20.56%, and 13.06% correspond to phase noise factors of 0.47, 0.45, and 0.46, respectively. That is, the phase noise factors for the reconstructed speech are: 0.47 at -10 dB, 0.45 at -5 dB, and 0.46 at 0 dB. We can verify the correctness of these results by substituting them into (5). So on average the resulting phase noise factor after phase reconstruction is 0.46.

VII. Comparison of the Results for the Three Cases of Phase Noise

To further analyze the test results obtained in the second set of experiments, we have calculated the average error rate difference (over all speakers) between the reconstructed phase and the perfect phase as well as the average error rate difference between the random phase and the perfect phase. Fig. 10 shows the results. As shown in this figure, the reconstructed phase case has a WER that is higher than that of the perfect phase by around 3.6% to 4.2% depending on the value of SNR. In other words, the reconstructed phase has a WER which is close to that of the perfect phase case. Furthermore, this difference is almost SNR independent. On the other hand, the WER difference between the random phase and the perfect phase ranges from 12.8% to 31.1%, which is much larger than those obtained in the reconstructed case. In this case, the value of the WER difference is

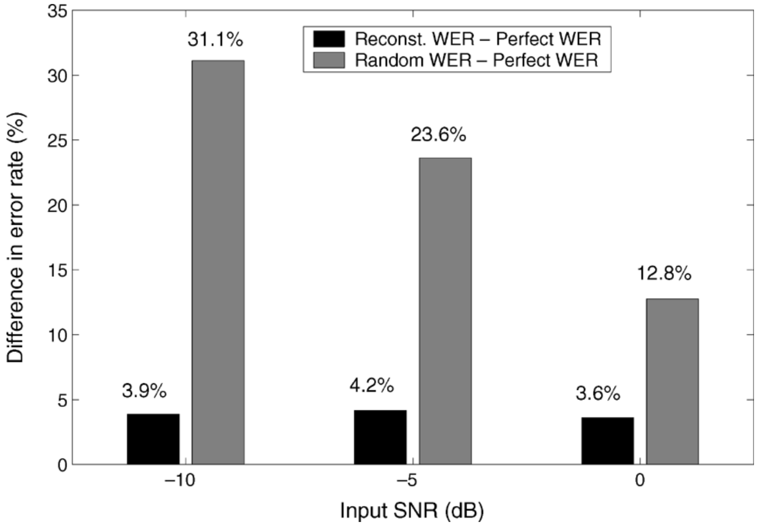


Fig. 10. Average error rate difference at different SNRs: Reconst. WER - Perfect WER = average error rate difference between the reconstructed phase and the perfect phase, Random WER - Perfect WER = average error rate difference between the random phase and the perfect phase.

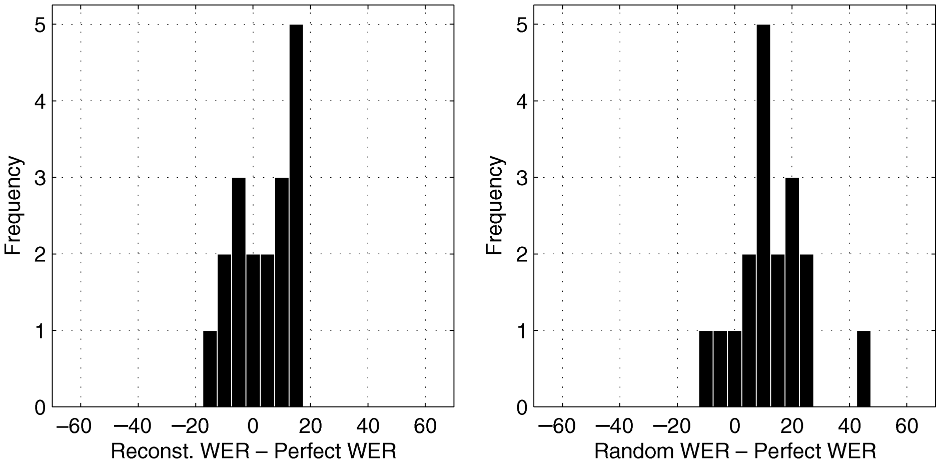


Fig. 11. Histogram of the WER difference at 0 dB. (Left: Reconstructed phase WER minus perfect phase WER. Right: Random phase WER minus perfect phase WER).

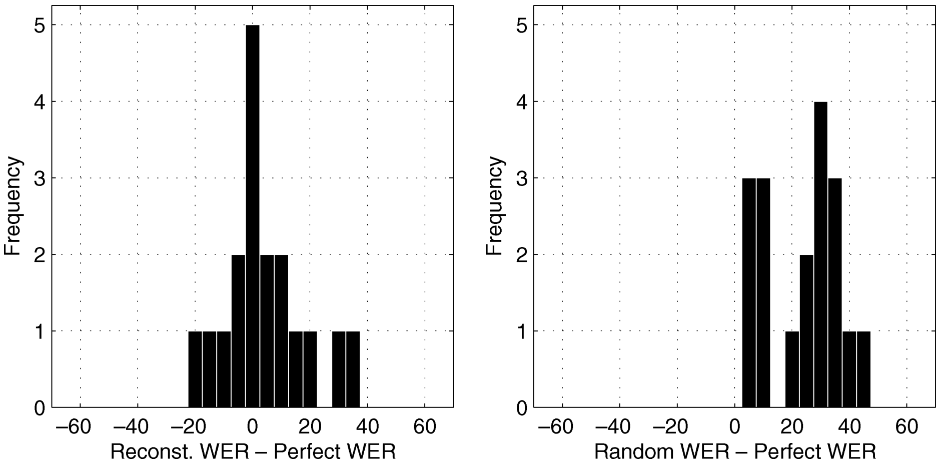


Fig. 12. Histogram of the WER difference at -5 dB. (Left: Reconstructed phase WER minus perfect phase WER. Right: Random phase WER minus perfect phase WER).

very much dependent on the SNR and increases as the SNR decreases. This shows that as the SNR decreases the importance of having a correct phase for recognition of speech increases.

To investigate this matter further, we have plotted the histograms of the WER difference for the cases of reconstructed phase and random phase at 0, -5, and -10 dB in Figs. 11–13,

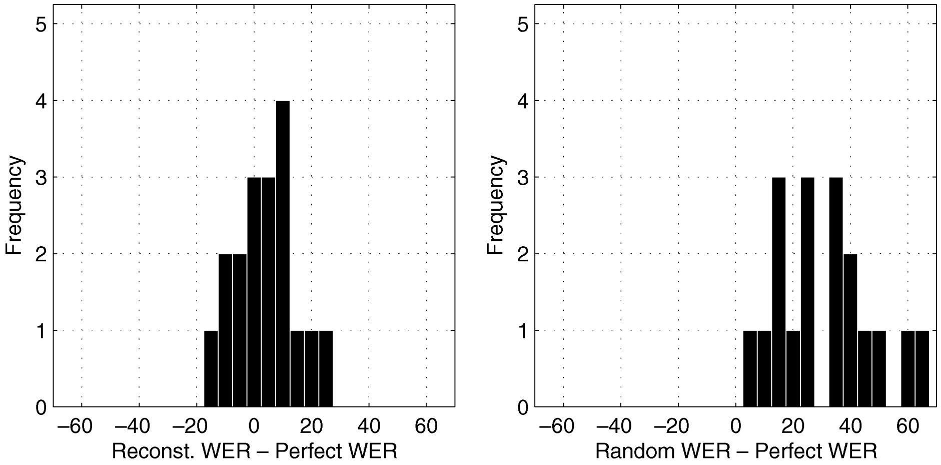


Fig. 13. Histogram of the WER difference at -10 dB. (Left: Reconstructed phase WER minus perfect phase WER. Right: Random phase WER minus perfect phase WER).

respectively. In each figure, the plot on the left corresponds to the WER difference between the reconstructed phase and the perfect phase; the plot on the right corresponds to the WER difference between the random phase and the perfect phase. A negative WER difference indicates that the word error rate for the case of the reconstructed phase or random phase is less than the perfect phase. Once again, we can see from these plots that the perception of phase varies among different listeners. For the reconstructed phase cases, a small number of people have a negative WER difference. This could be attributed to the close performance of the perfect phase and reconstructed phase or the limited range of the vocabulary used. For the random phase cases, all listeners performed considerably better in the perfect phase case when the SNR is -5 dB or -10 dB. When the SNR is 0 dB, about 10% of the listeners have a slightly negative WER difference which again shows that at higher SNR values phase is not as important for recognition of speech as it is at lower SNR values. In all three figures, we observe that using the reconstructed phase, the percent error difference has been shifted toward 0. This shows that much of the phase information can be recovered through the LSEE based phase reconstruction algorithm.

VIII. Conclusion

In this paper, we have investigated the relation between uncertainty in phase and recognition error rate of human listeners. Experimental results on 18 listeners (each attempting to recognize 360 words with different phase noise factors and SNRs) show that the effect of phase varies with SNR. At high SNRs (such as 20 dB) the effect of phase on the recognition error rate is small despite the fact that one can still perceive the existence of phase noise. At low SNRs (such as 0, -5, and -10 dB), the effect of phase on the recognition error rate can be significant. In such cases, the recognition error rate is more sensitive to phase noise when the phase noise factor is between 0.4 and 0.6. When the phase noise factor is less than 0.4, it has no significant effect on recognition rate. On the other hand, when the phase noise factor is greater than 0.6, there is a consistent and significant increase of error rate. In general the recognition error rate seems to be a sigmoidal function of the phase noise factor. We have also showed through experiments that LSEE based phase reconstruction can yield good results. The WER difference in recognition rate between the reconstructed phase case and the perfect phase case is about 4% on average.

There are a number of avenues for further exploration. In this paper, we have only considered the effect of phase when speech signals are corrupted by Gaussian noise. Further experiments on other noise types such as speech noise from competing speakers and reverberation could be useful. The experimental results in this paper show that the effect of phase on recognition rate varies among different listeners. Further experiments on variability among different words for a single speaker is worth studying. It should be mentioned that although phase information is not utilized in most speech recognition systems, phase information plays a crucial role in microphone array based applications such as time delay estimation and speech enhancement.

Acknowledgment

The authors would like to thank A. Keshavarz for her help in conducting the speech recognition experiments and the reviewers who provided constructive comments for this paper.

References

[1] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 3, Beijing, China, Oct. 2000, pp. 806–809.

[2] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[3] G. Shi and P. Aarabi, "Robust digit recognition using phase-dependent time-frequency masking," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Hong Kong, Apr. 2003, pp. 684–687.

[4] A. C. Lindgren, M. T. Johnson, and R. J. Povinelli, "Speech recognition using reconstructed phase space features," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Hong ++Kong, Apr. 2003, pp. 60–63.

[5] R. Schlüter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Salt Lake City, UT, May 2001, pp. 133–136.

[6] M. S. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Munich, Germany, Apr. 1997, pp. 375–378.

[7] R. J. McAulay and T. F. Quatieri, “Sinusoidal coding,” in *Speech Coding and Synthesis*. New York: Elsevier, 1995, pp. 121–173.

[8] H. Pobloth and W. B. Kleijn, “On phase perception in speech,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Phoenix, AZ, Mar. 1999, pp. 29–32.

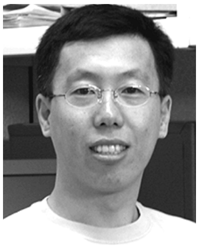
[9] D. S. Kim, “Perceptual phase quantization of speech,” *IEEE Trans. Speech Audio Processing*, vol. 11, no. 4, pp. 355–364, Jul. 2003.

[10] L. Liu, J. He, and G. Palm, “Effect of phase on the perception of intervocalic stop consonants,” *Speech Commun.*, vol. 22, no. 4, pp. 403–417, 1997.

[11] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[12] K. K. Paliwal and B. S. Atal, “Frequency-related representation of speech,” in *Proc. Eur. Conf. Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sep. 2003, pp. 65–68.

[13] K. K. Paliwal and L. Alsteris, “Usefulness of phase spectrum in human speech perception,” in *Proc. Eur. Conf. Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sep. 2003, pp. 2117–2120.



Guangji Shi (S’03) received the B.A.Sc. degree in computer engineering from the University of Minnesota, Minneapolis, in 1996, the M.A.Sc. degree in electrical and computer engineering from the University of Toronto (UT), Toronto, ON, Canada, in 2002, and is currently pursuing the Ph.D. degree in electrical and computer engineering at UT.

Before joining UT, he worked in the automation industries both as a Technical Engineer and as a Software Developer. His current research interests include robust speech recognition, microphone arrays, and image processing. His research on phase-based dual-microphone speech enhancement has appeared in Scientific American.

Maryam Modir Shanechi joined the graduate program at the Massachusetts Institute of Technology (MIT), Cambridge, in 2004 after completing her bachelor’s degree in engineering science (electrical option) at the University of Toronto, Toronto, ON, Canada. Her undergraduate research included problems of speech separation, sound localization, and speech recognition.

Her current research interests lie in problems of communications, information theory, and signal processing and she is currently a member of the Signals, Information and Algorithms Laboratory at MIT. She held summer internships at Altera Corporation in 2004 and Vanu, Inc. in 2005.

Ms. Shanechi has received a number of awards for academic achievement, including the Professional Engineers of Ontario (PEO) gold medal, the Wilson Medal, and NSERC scholarships. She has been selected by the *University of Toronto Magazine* as one of the next generation of Canadian leaders.

Parham Aarabi (S’97–M’01) received the Ph.D. degree in 2001 in electrical engineering from Stanford University, Stanford, CA, the M.A.Sc. degree in 1999 in computer engineering from the University of Toronto, Toronto, ON, Canada, and the B.A.Sc. degree in 1998 in engineering science (electrical option) from the University of Toronto.

H is a Canada Research Chair in Multi-Sensor Information Systems, a tenured Associate Professor in the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, and Founder and Director of the Artificial Perception Laboratory, all at the University of Toronto. His current research, which includes multisensor information fusion, human–computer interactions, and VLSI implementation of sensor fusion algorithms, has appeared in over 50 peer-reviewed publications and covered by media such as the *New York Times*, *MIT’s Technology Review Magazine*, *Scientific American*, *Popular Mechanics*, the Discovery Channel, CBC Newsworld, Tech TV, and City TV.

Dr. Aarabi received the 2002, 2003, and 2004 Professor of the Year Awards, the 2003 Faculty of Engineering Early Career Teaching Award, the 2004 IEEE Mac Van Valkenburg Early Career Teaching Award, and the 2005 Gordon Slemon Award.