PERCEPTION OF STOP CONSONANTS WITH CONFLICTING PHASE AND MAGNITUDE

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

ABSTRACT - Identification experiments were performed to examine the relative importance of phase versus magnitude for stop consonant perception. Three types of stimuli were constructed from Vowel-Consonant-Vowel (VCV) utterances: (1) Swapped stimuli, a swapped stimulus has the magnitude spectra of its consisting patches from one VCV signal and its corresponding phase spectra from another; (2) Phase-only stimuli, a phase-only stimulus is constructed by eliminating the original signal's magnitude information; (3) Magnitude-only stimuli; It was shown that the perception to the intervocalic stop consonant in a swapped stimulus varied from magnitude dominance to phase dominance as the size of analysis patches increases. The crossover lies somewhere between 192 ms and 256 ms where both magnitude and phase spectra provide equivalent but conflicting perceptual information. It was found that the perception of voicing property (voiced/voiceless) of a stop consonant rely strongly on its phase information while the perception of a stop's place of articulation was mainly determined by its magnitude information. As the patch size increases from 16 ms to 512 ms, the identification rates to the intervocalic consonants in magnitude-only stimuli decrease from 78% to 30% and in phase-only stimuli increase from 18% to 68%.

INTRODUCTION

The perception of acoustic signals applied to the ear is largely determined by the signals' magnitude spectra while their phase spectra are generally accepted as less or unimportant. The fact that considerable phase distortions do not bring any noticeable effect on the perception of many musical notes and synthesized vowels support this opinion. Helmholtz's conclusion that "the quality of the musical portion of a compound tone depends solely on the number and relative strength of its partial simple tones, and in no respect on their differences of phase" (von Helmholtz, 1954, p. 126) had led many investigators for the greater part of a century to believe that the ear's sensitivity to phase could be regarded as negligible, in another word, the auditory system was phase deaf. In many subsequent studies, however, certain phase sensitivity were noticed, particularly for phase patterns which result in markedly different time waveform envelopes. Most of the studies on the perceptual influence of phase were carried out using stimuli containing several frequencies or a number of harmonics of a single frequency (Chapin and Firestone, 1934; Mathes and Miller, 1947; Zwicker, 1952; Licklider, 1957; Schroeder, 1959; Goldstein, 1967; Buunen *et al.* 1974; Patterson, 1987). A noticeable shift in interest has taken place in psychoacoustics during the past twenty years away from simple and static sounds towards more complex and dynamic sounds. The main moving forces behind this shift of interest may have been, among others, the growing awareness that the auditory system as a whole is so non-linear and time-variant that its total behavior will never be completely understood in terms of its behavior for simple steady-state sinusoidal tones. Phase effects arising from a complex signal such as speech are not easy to predict, if not impossible, from experiments using simple stimuli.

Despite the general realization of audibility of phase changes, the effects of phase on speech perception have been received far less attention than the effects of other acoustic cues. Vowels generated by adding sine-waves of appropriate magnitude and phase have been used predominantly in the studies of phase influences on vowel or vowel-like perceptual qualities. The experiments made by Schroeder and Strube (1986) using stimuli consisting of 31 equal-magnitude harmonics revealed that many realizable timbres have vowel-like perceptual qualities when the fundamental frequency lies in the human pitch range, suggesting the possibility of constructing intelligible voiced speech signals that have flat-magnitude spectra. The observed vowel timbres were explained by the dependence of the short-time magnitude spectra on phase changes. Darwin and Gardner (1986) made observations of phase effect on vowel quality using a series of synthetic vowels with different first formants (F1s) under three kinds of phase pattern. They generated the vowels using additive sine-wave synthesis based on Klatt's (1980) cascade synthesizer and discovered that mistuning of only 1-2 Hz in harmonics close to F1 caused a lowering of the phoneme boundary in a categorization task. Traunmüller (1987) found that phase information is sufficient to convey the information necessary to

distinguish vowels if the fundamental frequency F0 is low enough. A study by Oppenheim and Lim (1981) demonstrated that speech can still be recognized in a Fourier reconstruction despite defective magnitude information. They even noticed that the long-time phase information prevails over magnitude. For example, if the magnitude and phase spectra of two speech signals are interchanged, the resulting hybrid signal is recognizably that from which the phase spectrum is taken. The modifier "long-time" in the above formulation is necessary because the prevalence of phase over magnitude demonstrated by Oppenheim and Lim was obtained on a 2 seconds time base. More recently, the phase effect on the perception of intervocalic stop consonants has been studied by Liu *et al.* (1992). Their study showed that on a time base of 0.5 second the perception of stop consonants depends more on phase than on magnitude.

METHOD

The present study is concerned with a further investigation on the importance of phase information for consonant perception. In discrete Fourier transform, a time signal $x(n)$ can be represented as:

$$X(k) = A_k e^{j\phi_k} = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2p}{N}kn} \qquad k = 0, 1, \cdots, N-1$$

the signal's Fourier magnitude spectrum is given by $\{A_0, A_1, \cdots, A_{N-1}\}$ and its phase spectrum by $\{\phi_0, \phi_1, \cdots, \phi_{N-1}\}$. In a patchwise Fourier analysis the speech signal is divided into a number of overlapped patches. From each patch we can obtain its phase and magnitude spectrum. It is clear that the acoustic information resides in both phase and magnitude spectra of all patches. A comparison of the relative importance of the two kinds of information for perception of natural VCV utterances was first conducted by using a set of stimuli which had its magnitude from one VCV signal and its phase from another. That is, the magnitude and phase information in each of these reconstructed stimuli are perceptually conflicting. We denote a stimulus of this kind as $VC_AV \circ VC_\phi V$, having its magnitude from VC_AV and its phase from $VC_\phi V$. We refer to C_A as the stimulus' *magnitude-consonant* and C_ϕ as its *phase-consonant*. There are two principles in the synthesis: (1) C_A and C_ϕ are different stop consonants; (2) The two utterances which provide respectively the magnitude and phase to a stimulus come from the same speaker. Figure 1 shows schematically the procedure used for this phase-magnitude swapping between two VCVs.

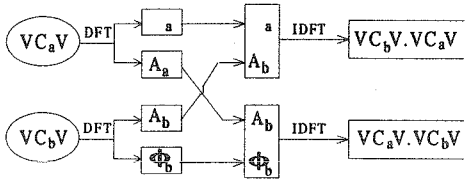

Figure 1. VC_aV and VC_bV are the original speech signals. These are subjected to DFT. The phase ϕ_a, ϕ_b and magnitude A_a, A_b are cross-combined and subjected to IDFT to yield two stimuli $VC_bV \circ VC_aV$ and $VC_aV \circ VC_bV$.

Phase-only and magnitude-only stimuli were also used. Phase-only or magnitude-only stimulus can be constructed from each speech signal by eliminating the original VCV's magnitude information or its phase information. Such stimuli used in our experiments were constructed by setting each patch's original magnitude or phase to a constant value.

EXPERIMENT

A. VCV Recording

Vowel-Consonant-Vowel (VCV) syllables formed by pairing each of the six stop consonants /p, b, t, d, k, g/ with vowel /a/ were spoken by 10 untrained German speakers (6 male, 4 female). The speakers

were asked to read each VCV in isolation twice with approximately equal stress on the two syllables. The signals were low-pass filtered with a cut-off frequency of 7.5 kHz and the sampling rate was 16 kHz. The duration of each recorded speech signals was in the range of 300-500 ms. All signals were then expanded to a duration of 512 ms (8192 points) by appending silent periods before and after the recorded syllables. VCVs spoken by the same speaker were manually aligned according to their CV portions.

B. Subjects

13 paid students (8 male, 5 female, age 20-35) served as subjects. All were native German speakers with no known history of either speech or hearing disorder. They had no previous experience in psychoacoustical experiments. None of them served as a speaker.

C. Stimuli

Seven different window sizes were used in our experiments: 16, 32, 64, 128, 192, 256 and 512 ms. The advancing step of the analysis window was half of its size. The stimuli were reconstructed from their short-time spectra by overlap addition method (Rabiner and Schafer 1978). From each speaker we had two versions of 210 $VC_AV \circ VC_\phi V$ stimuli (30 consonants-combinations x 7 window sizes). One version was used for practice and the other for test. Stimuli resynthesised using non-conflicting magnitude and phase spectra (swapping between 2 versions of each VCV syllables) were also included as a control for the identifiability of the swapped stimuli. There were 12 such control stimuli from each speaker. The sequence of a total of 2220 testing stimuli from the 10 speakers was randomized. It was then split up into 6 testing blocks, each containing 370 swapped stimuli. From 840 phase-only and magnitude-only stimuli (6 stops x 10 speakers x 7 analysis windows x 2) another two randomized testing blocks were prepared. Stimuli were played back through a YAMAHA AX-930 sound amplifier.

D. Procedure

The subjects were tested in small groups (2-4 subjects) sitting in a sound-isolated booth. All stimuli were presented binaurally over headphones at a comfortable listening level. The subjects were instructed to make a forced-choice identification to the intervocalic consonant and write their responses on response sheets. On the top of each response sheet the six letters /p, b, t, d, k, g/ from which a response should be selected were displayed. The subjects were first familiarized with the task by listening to a 150-stimuli training block. The interstimulus interval was 1.2 second. Each testing block lasted about 12 minutes and there was a 5-10 minute break after each block. The experiments were conducted in two sessions. Each session consisted of four stimuli blocks. For each subject the second session was performed several days later after the first session.

RESULT AND DISCUSSION

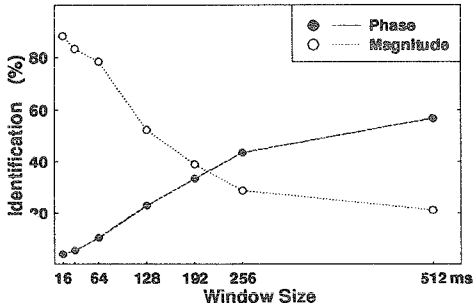

Figure 2. Identification results for $VC_AV \circ VC_\phi V$ stimuli.

The percentages of C_A and C_ϕ responses to the hybrid $VC_AV \circ VC_\phi V$ stimuli are shown in figure 2. The percent of C_A identification decreases monotonically while that of C_ϕ identification increases monotonically with the size of analysis window. in other words, the identification to the intervocalic stop consonant is dominated by magnitude for smaller analysis windows while phase becomes more important when the analysis window is getting larger. There is a crossover from magnitude dominance to phase dominance at a window size somewhere between 192 ms and 256 ms where the phase and magnitude provide conflicting but equivalent perceptual information. Note that the sum of the two identification rates at each window size is not 100% since our experiments were based on six alternative choices. The rate of C_ϕ responses at the window size of 512 ms was 58% which is in good agreement with data reported previously (Liu *et al.*, 1992).

In order to evaluate the relative importance of phase versus magnitude with respect to the voicing and place properties of the composing stop consonants C_A and C_ϕ, the identification results are replotted in figure 3 after dividing the 30 kinds of C_A and C_ϕ combinations into three groups: (1) C_A and C_ϕ have the same voicing property (both voiced or both voiceless) but are different in their place of articulation. There are 12 such combinations, i.e., {p/t, p/k, t/p, t/k, k/p, k/t, b/d, b/g, d/b, d/g, g/b, g/d}; (2) C_A and C_ϕ have the same place of articulation (either labial, alveolar or velar stops) while different in their voicing property. Six pairs are members of this group: {p/b, b/p, t/d, d/t, k/g, g/k}; (3) the third group is composed of the rest 12 C_A and C_ϕ combinations in which the two stops in each pair differ not only in their voicing property but also in their place of articulation. The data of the left panel in figure 3 are based on stimuli of the first group and the data of the right panel from the second group. More exactly, figure 3(a) shows the results regarding to the perception of stops' place of articulation and figure 3(b) shows the results regarding to the perception of stops' voicing properties. The proportion of C_A responses under the same window size is clearly higher in figure 3(a) than that in figure 3(b). This means that a C_A response is more likely when C_A and C_ϕ of a swapped stimulus are of different place of articulation than when they have different voicing properties. This conclusion held for all 7 window sizes.

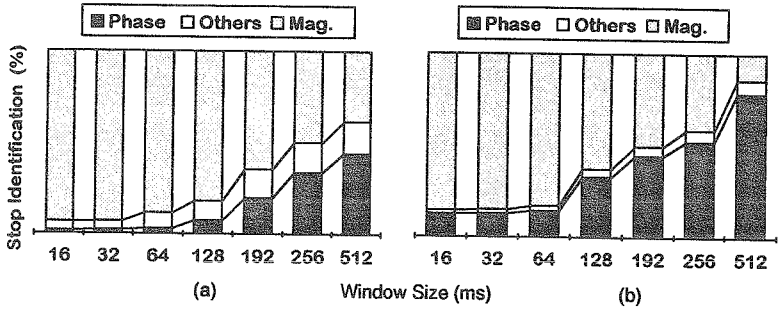

Figure 3. Identification results for $VC_AV \circ VC_\phi V$ stimuli. (a) responses by grouping stimuli according to their voice/voiceless properties. (b) responses by grouping stimuli according to their place of articulation.

By examining figure 4 carefully, we found that magnitude and phase information tend to play different roles in the perception of stop consonants. As magnitude information appear to be more responsible for place identification, phase, in the meanwhile, determines whether the stop is voiced or unvoiced. In the case of a stimulus' composing C_A and C_ϕ having different place of articulation and different voicing, it is more likely that this stimulus is perceived as the stop consonant which share the place of articulation with C_A but the voicing with C_ϕ. For example if the stimulus' C_A and C_ϕ are /b/ and /t/ respectively, it is more likely that this stimulus is perceived as /p/. There are 12 such $C_A\_C_\phi$ combinations in our swapped-stimuli. In figure 4 the 130 responses (10 speakers x 13 subjects) under each $C_A\_C_\phi$ combination were presented by 6 bars connected together bottom-up in the order of P-B-T-D-K-G. Dark bars in the figure

represent the responses for 3 voiced stops (B-D-G) and light bars for the 3 unvoiced stops (P-T-K). It can be seen that the most likely to be perceived consonant in each case is C_A, followed by the one which share the place of articulation with C_A but the voicing with C_ϕ.

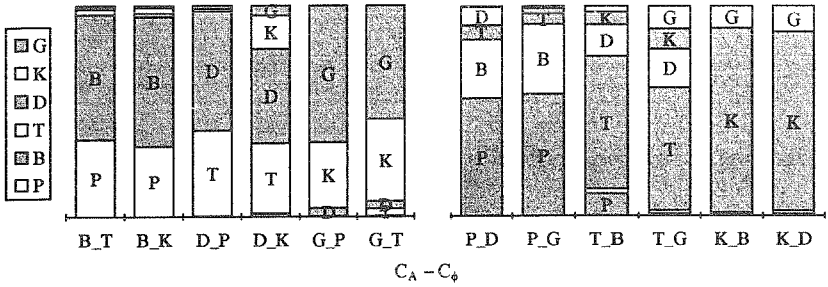

Figure 4. Identification results based on $VC_AV \circ VC_\phi V$ stimuli from the third group. The window size 128 ms.

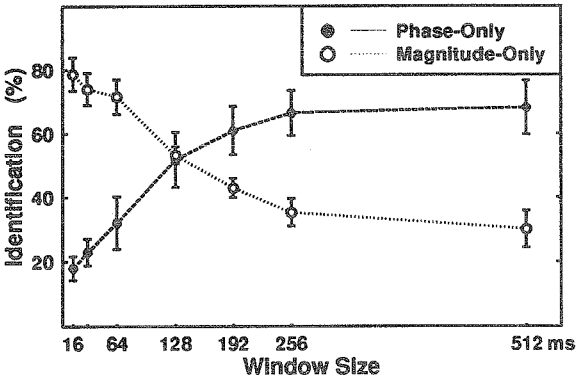

Figure 5. Identification results to phase-only and magnitude-only stimuli.

The use of phase-only and magnitude-only stimuli provide yet another way of comparing the effect of a stimulus' magnitude spectrum with that of its phase spectrum on the perception of stop consonants. The two curves in figure 5 present the average identification performance and the standard deviation as a function of window size for magnitude-only and phase-only stimuli. At the window size of 128 ms the identification rates for the intervocalic stop consonants from two kinds of stimuli are approximately equal.

In summary, we found that (1) the relative importance of phase versus magnitude for the perception of intervocalic stop consonants changes with the analysis window size. When the window size is small the magnitude spectrum is perceptually more important than the phase spectrum. However, as the analysis window size is getting larger the identification of the intervocalic stop consonant will depend more and more on its phase spectrum. The transition from magnitude dominance to phase dominance is at a window size somewhere between 192 ms and 256 ms; (2) a stop consonant's voicing characteristic rely strongly on its phase information, on the other hand, the place characteristic is mainly determined by its magnitude information, especially in the case of smaller analysis windows.

REFERENCE

Buunen, T.J.F., Festen, J.M., Bilsen, F.A., and van den Brink, G. (1974). "Phase effects in a three-component signal," J. Acoust. Soc. Am. 55, 297-303.

Chapin, E.K., and Firestone, F.A. (1934). "The influence of phase on tone quality and loudness; the interference of subjective harmonics," J. Acoust. Soc. Am. 5, 173-180.

Darwin, C.J., Gardner, R.B. (1986). "Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality," J. Acoust. Soc. Am. 79, 838-845.

Goldstein, J.L. (1967). "Auditory spectral filtering and monaural phase perception," J. Acoust. Soc. Am. 41, 458-479.

Von Helmholtz, H.L.F. (1954). "On the Sensations of Tone," Dover, New York. (Originally published, 1885).

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. 67, 971-995.

Licklider, J.C.R. (1957). "Effects of changes in the phase pattern upon the sound of a 16-harmonics tone," J. Acoust. Soc. Am. 29, 780.

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, 2340.

Mathes, R.C., and Miller, R.L. (1947). "Phase effects in monaural perception," J. Acoust. Soc. Am. 19, 780-797.

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

Patterson, R.D. (1987). "A pulse ribbon model of monaural phase perception," J. Acoust. Soc. Am. 82, 1560-1586.

Rabiner, L. R. and Schafer, R. W. (1978). "Digital processing of speech signals," Prentice-Hall, Inc. New Jersey.

Schroeder, M.R. (1959). "New results concerning monaural phase sensitivity," J. Acoust. Soc. Am. 31, 1579.

Schroeder, M.R. and Strube, H.W. (1986). "Flat-spectrum speech," J. Acoust. Soc. Am. 79, 1580-1583.

Traunmüller, H. (1987). "Phase vowels," in: Schouten, M.E.H.: *The psychophysics of speech perception.* Martinus Nijhoff, Dordrecht, pp. 377-384.

Zwicker, E. (1952). "Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones," Acustica 2, 125-133.