THE EFFECTS OF PHASE INFORMATION ON THE INTELLIGIBILITY OF CHANNEL VOCODED SPEECH

R.H. Mannell
Speech, Hearing and Language Research Centre
Macquarie University

ABSTRACT - The intelligibility of vocoded speech with various phase spectra was compared to the intelligibility of the original input natural speech. It was found that vocoded speech with true natural phase was the closest to natural speech in intelligibility.

INTRODUCTION

Hermann Helmholtz, in the mid to late 1800's, logically extended Ohm's Acoustic Law by observing that the ear is effectively insensitive to phase. Helmholtz, however, "confined his conclusion to the 'musical' portion of the sound" (Wever 1949, p419). Wever outlines various early studies of phase perception which generally support Helmholtz's notion when confined to periodic signals with steady state phase and amplitude characteristics. Some of the studies he reviewed, however, indicated that rapidly changing phase can be perceived due to the effects of phase relationships (reinforcement and cancellation) on signal amplitude. It is well known that the shape of the time domain waveform of a complex signal is very sensitive to changes in the phase relationships of its various frequency components. Figures 3B to 3F show how greatly the waveform of the vowel /i/ varies with changing phase. This variation occurs without any significant change in the intelligibility of the vowel. This well known perceptual insensitivity to phase in vowels has been used as one of the basic assumptions of channel vocoder techniques. Channel vocoders typically consist of a bank of B.P. filters with linear zero phase spectra. The omission of phase from the transmitted signal is one of the major reasons for the moderate transmission bandwidth savings available when using this type of vocoder.

Gold (1964) cited two earlier studies which indicated that "there is some evidence that severe phase distortion introduced by both pitch-excited and voice-excited vocoders causes deterioration in the quality of the synthetic speech" (ibid. p1892). Gold's study itself produced an unintelligible wave derived from three simple formant trackers, but which had "speechlike phase" and used this wave to excite the vocoder synthesiser (rather than the usual zero phase buzz and random phase hiss). Gold found that his listeners (in informal listening tests) reported that the output speech sounded natural and no longer possessed the then typical vocoder quality. Flanagan and Golden (1966) extracted both phase and amplitude information from natural speech in their "phase vocoder" and recombined the two signals on resynthesis, thus avoiding the need for separate voicing and pitch analysis and excitation. In this system, both phase and amplitude were band limited and transmitted (rather than just the amplitude information as in normal channel vocoders). The main difficulty with this system was the need to produce a differential phase spectrum which could be band-limited (unlike the normal phase spectrum, which is unbounded). The bandlimited differential phase values could then be transmitted (along with the bandlimited amplitude values) and the phase could be restored by integration before recombining with amplitude at resynthesis. The synthetic speech produced by this system was claimed to "considerably surpass[]" the quality of normal channel vocoders.

Oppenheim et al (1979) demonstrated that speech with the amplitude spectrum set to unity and the phase spectrum retained intact produced "phase-only" speech with unnatural (noisy) quality but with a high degree of intelligibility. Spectrograms of this speech showed that the formant structure of the speech had been maintained. These results parallel similar findings with phase in image reconstruction. Oppenheim (1981) further demonstrated that speech with phase set to zero (amplitude maintained) is less intelligible than speech with amplitude set to unity (phase maintained). These results were obtained by manipulation of the long-time Fourier transforms of these signals. He concluded that phase in short-time spectra is insignificant. This is presumably because short-time spectra could be said to be modelling speech as a quasi-stationary signal in which the frequency components are approximately stationary. Further, in the case of long-time spectra, speech is no longer being modeled as a quasi-stationary signal, but a signal in which the frequency components change dynamically in time. Oppenheim concludes that for both speech and images phase information preserves the "location of

events" such as "lines edges and other narrow events" (ibid. p534). In other words, the long-time phase spectrum encodes the location of major changes in the signal, and in speech this implies changes in amplitude of frequency components. It is not surprising, therefore, that "phase-only" speech is reasonably intelligible as normal continuous speech is continuously changing and it is these changes that are preserved by the long-time phase spectrum. The long-time amplitude spectrum only encodes the average values for each of the frequency components and so effectively time-smears the signal when the phase information is omitted.

Even though channel vocoding might be considered to be a short-time analysis and as such, the phase spectrum should therefore be unimportant, there are several phonemes which involve rapid changes in amplitude (associated with rapid opening, closing, coupling or uncoupling of resonator chambers). It is reasonable to assume that such rapid changes might constitute important perceptual cues. Further, it is possible that at least some of these cues might require precise identification of their location (analogous to edge location in images) or of the shape of their time-domain waveform amplitude envelope. It is exactly these phonemes for which the short-time fourier analysis is able to least accurately model as a quasi-stationary system. It is therefore likely that it is these phonemes which most require the edge location information supplied by the phase spectrum and these phonemes which will suffer most from inadequate phase information. Such consonants include the stops and affricates (opening or closing of the oral cavity), the nasal consonants (coupling or decoupling of the nasal cavity and opening or closing of the oral cavity), and the lateral /l/ (coupling or decoupling of the two parallel oral cavities). This principle could also apply to the fricatives as their constrictions effectively decouple the vocal tract posterior to the constriction.

The relative inertia of the various articulators cause the openings and closings of the oral cavity to proceed at different rates for the different places of articulation. Presumably the time domain waveform amplitude envelope of these different places of articulation may not only differ in their rates of amplitude change but also perhaps in their overall shapes. Thus we would expect labial and velar articulations to move more slowly than alveolar articulations which involve the fast moving tongue tip. Thus, /t/, /d/ and /l/ would exhibit faster changing steeper amplitude envelopes at the point of release. It is essential, at this point, to exclude from consideration those phonemes which have strong spectral cues. This would include the alveolar and post-alveolar fricatives and /t/ and /k/ which typically have strong bursts and aspiration in CV position in Australian English. /p/ on the other hand tends to be weakly aspirated as do the voiced stops and so the waveform envelopes might compete with formant transitions as cues to phoneme identity. The remaining phonemes likely to be affected by inaccurate edge detection or waveform envelope determination are the labial stops, the other voiced stops, the weak labial and dental fricatives and /l/. /w/ is a possible addition to this list. /j/ and /r/ on the other hand have very distinct formant transition cues which are likely to predominate over waveform envelope cues.

EXPERIMENT

This experiment was carried out utilising the channel vocoder described in Clark and Mannell (1988). Natural speech intelligibility was compared with the intelligibility of speech output by five vocoder configurations. In all five cases the vocoders had auditorily scaled filter banks with all filters uniformly 1 Bark in bandwidth. The only difference between the five vocoders was the treatment of phase.

One vocoder ("ZERO") was a typical channel vocoder in that all of its filters had linear zero phase.

A second set of data ("NATPHAS") was derived from the output of this same vocoder but its synthetic phase spectrum was replaced with the original natural phase spectrum. The synthetic amplitude spectrum and the natural phase spectrum were used to extract new real and imaginary spectral components which were then inverse FFT'ed to obtain a composite waveform. This was possible as the natural input speech and the synthetic output speech of ZERO were precisely time aligned. In figure 2, D represents the speech output by ZERO whilst B represents that same speech with natural phase restored (NATPHAS). With the exception of the extra high frequency components caused by the width of the higher frequency filters, the waveform produced by the condition NATPHAS is very similar to that in the original natural signal (2A).

The third vocoder condition ("DELAY") was passed through identical filters to ZERO, however upon remodulation of the synthetic source with the channel amplitude information the impulses of the voiced source were delayed by one sample for each filter with increasing centre frequency so that the 18th filter was delayed by 18 samples or 1.8 msecs. This progressive delay is equivalent to an increasingly negative phase delay as frequency increases. This arrangement conforms approximately to Fant's

requirement that "if the phase shift is not linearly related to frequency, there will be separate time delays for separate frequency intervals of the spectrum" (Fant, 1960, p235). Note that only the voiced components of the spectrum can be encoded for phase in this way. This is reasonable, however, since in "...random white noise ... phase is distributed at random throughout the spectrum" (ibid, p235). The waveform shape for the vowel /i/ is shown in figure 3C and could be claimed to be closer to A and B than are any of the vocoder outputs shown in 2D, 2E or 2F.

The fourth vocoder condition utilised the same filter as before except that their phase was changed. The phase values were chosen on the basis of a statement by Fant (1968) to the effect that "...the phase shifts by the amount $-\pi$ per formant with increasing frequency" (ibid, p195). The theoretical phase spectrum of the 11 Australian English monophthongs (calculated according to this formula) are shown in figure 2 as dotted lines. The solid line represents the phase spectrum of a neutral vowel with formants at odd multiples of 500 Hz. These values would be the ideal choice for the filter phase values, however, such values are restricted to the range 0° to -360° and so it was necessary to wrap the ideal phase spectrum into this range to give the dashed line in figure 2. The arrows in this figure represent the filter centre frequencies and the phase value given to each of these filters corresponds to the values of the dashed line immediately below each arrow. The waveform of /i/ for this vocoder is shown in figure 2E.

The fifth vocoder was produced in a similar way to vocoder 4 except that the phase spectrum declined linearly from 0° at 0 Hz to -360° at 5 kHz. The waveform of /i/ for this vocoder is shown in figure 2F.

The test tokens represented most of the orthographically unambiguous vowels (11) and consonants (19) of Australian English presented in /h_d/ and CV (V=/a/) frames respectively. They were recorded by a male speaker in a sound treated room to professional audio standards. The use of nonsense tokens ensured that the listeners made minimal use of linguistic context in making their judgments.

Twenty subjects were screened to ensure that they were able to identify similar speech tokens at a presentation level at least 30 dB s.p.l. below that of the actual test materials. Half the subjects were presented the (internally randomised) conditions in one order and the other 10 subjects were presented the conditions in reverse order (no differences were found for the two presentation orders). The tests were conducted in a sound treated room using TDH49 headphones with supra-aural cushions and circumaural seals at a presentation level of 70 dB s.p.l. (ref. 20 μP).

RESULTS AND DISCUSSION

The vocoder conditions ZERO, DELAY, and the two decreasing phase conditions all sound similar and possessed to some degree the faint buzziness typical of channel vocoders. The NATPHAS case on the other hand sounded more natural although it had a reverberant quality.

The listener responses are summarised in tables 1 and 2. All values are percentage correct responses for each phonetic category or phoneme. The symbols represent the results of Chi square tests. The underlined symbols represent significant deviation from the reference case with p.01, whilst the symbols without underlining represent p.05.

An important general trend is evident. The two decreasing phase vocoders are generally significantly less intelligible than the other cases. This suggests that either the assumptions about the need for decreasing phase are incorrect or (more likely) that this method of encoding decreasing phase is not to be recommended. Clearly the 0° to -810° case is not accurately modeled and can only be modeled accurately with a combination of filter phase (as used here) plus delays to model the unwrapping of the phase range. The 0° to -360° condition is clearly a bad model of natural phase spectrum. This suggests that bad phase modelling is worse than zero phase.

A second important trend is that the DELAY condition is generally not significantly different to the NATPHAS case. This is true for all phonemes and phonetic categories at the 1% level. At the 5% level the vowel category (but no individual vowels) and the consonant /t/ are significantly lower in intelligibility for condition DELAY relative to condition NATPHAS. This suggests that the delay method is a perceptually acceptable method of simulating natural phase.

There are two cases where ZERO phase has significantly lower intelligibility than NATPHAS (/v/ and /l/). In these two examples NATPHAS and DELAY are not significantly different to the natural intelligibility whilst ZERO is. (nb. the intelligibilities of these two phonemes are responsible for similar statistics for the voiced fricative and approximant classes). ZERO is never significantly more intelligible

than NATPHAS or DELAY although it is often superior to the two decreasing phase cases. The zero phase condition is therefore an acceptable approach for most phonemes but is less than optimal for a small number of phonemes.

For three of the stops, all vocoders produce intelligibilities significantly less than for the natural condition. It must be concluded that some cause other than phase is responsible for this deterioration in intelligibility.

It is useful to examine the confusion patterns produced by these vocoders. In the case of /p/, most vocoders produced sounds which were most often misidentified as /h/. The only exception is the DELAY vocoder where /p/ is more often heard as another voiceless stop. /b/ misidentifications are exclusively /v/ for ZERO and NATPHAS and almost so for DELAY but there is an increasing number of /dh/ (voiceless labiodental fricative) identifications as phase modelling deteriorates (through the 0° to -810° condition to the 0° to -360° condition). /d/ confusions are generally to /g/ (especially for the ZERO condition) but the number of voiced fricative (/dh/ and to a lesser extent /v/) confusions increases from NATPHAS to DELAY to the decreasing phase conditions. In other words increasingly poor performance is accompanied by more fricative identifications. /v/ confusions are nearly always with labiodental fricatives, and especially the voiced /dh/. The /l/ confusions are exclusively with /dh/ for ZERO with /z/ and /v/ appearing for the 0° to -810° condition and /w/ predominating for the 0° to -360° condition. Finally, the significant /w/ confusions in the 0° to -360° condition are all with /r/. The general trend in all this seems to be the increasing number of /dh/ mis-identifications as the phase information deteriorates. It is difficult to interpret this as a trend towards increased place or manner of articulation confusions. It may be that the trend is towards an increased number of labiodental place misidentifications and as the only labiodentals in English (in this CV context) are fricatives, there is a tendency for such identifications to cross manner boundaries only because there is no suitable place phoneme of the appropriate manner.

CONCLUSIONS

The NATPHAS condition is closest to natural speech in intelligibility although the insertion of natural phase did nothing to improve the intelligibility of the three low intelligibility stops. It is assumed that some other aspect of the frequency or temporal encoding is responsible for the low intelligibility of these sounds. The DELAY condition is closest to the NATPHAS condition in modelling natural phase. The ZERO phase condition is also close to the natural phase condition except for two consonants. These three conditions are all significantly higher in intelligibility than the other two conditions. This indicates that zero phase is to be preferred to "bad" phase. Modelling phase spectra by manipulating the phase spectra of the BP filters in a channel vocoder is not very successful. As it is impractical in a realistic vocoder system to reinsert natural phase after resynthesis, the best results appear to be achieved with a synthetic voice source impulse delay methodology. The exact value of the delay for each channel filter needs to be examined in more detail as the delay used in the present experiment was rather ad hoc and used because of its simplicity.

Most confusions occur in phonemes with rapid temporal changes in the waveform envelope. On the other hand, not all phonemes with rapid changes are confused. This is presumably because other cues predominate in the identification of these phonemes. Such cues might be formant transition cues or spectral envelope cues. It is likely that some alveolars (especially /l/ but perhaps also /d/) as well as the bilabials and labiodentals are not strongly differentiated from the apicodentals by formant transition and spectral envelope cues. In these cases it is possible that some form of phase-dependent waveform envelope cue might dominate the identification process. This could not occur for /t/ which has a strong aspiration spectrum, nor would it occur with the labial and alveolar nasals which have strong nasality cues.

REFERENCES

Clark, J.E. & Mannell R.H. (1988) "Some comparative characteristics of uniform and auditorily scaled channel synthesis", Proc. SST-88, 282-287.

Fant, G. (1960) Acoustic Theory of Speech Production, (Mouton: The Hague, second printing 1970)

Fant, G. (1968) "Analysis and synthesis of speech processes", in Malmberg, B. (ed.) Manual of Phonetics (North Holland: Amsterdam)

Flanagan, J.L. & Golden, R.M. (1966) "Phase vocoder", Bell Sys. Tech. J., 1493-1509.

Gold, B. (1964) "Experiment with speechlike phase in a spectrally flattened pitch-excited channel vocoder", J. Acoust. Soc. Am. 36, 1892-1894.

Oppenheim, A.V., Lim, J.S., Kopec, G. & Pohlig, S.C. (1979) "Phase in speech and pictures", IEEE, ICASSP 1979, 632-637.

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

Wever, E.G. (1949) Theory of Hearing (Dover Publications: New York, 1970 edition)

	NATURAL	Zero	NatPhas	Delay	0/-810	0/-360
p	100 #@	45*	45*	40*	50*	25*
t	95	100	100	100	100	100
k	100	100	100	100	100	100
b	95 #@	0*	5*	10*	0*	0*
d	100 #@	60*	35*	50*	5*#	10*#
g	100	90	100	100	80* @	100
tʃ	100	90	95	95	90	75* @
dʒ	100	100	100	100	100	100
f	100	100	100	80*#@	100	100
s	100	85	85	95	85	95
ʃ	100	100	95	95	100	95
v	100	80* @	100 #	90	75* @	80* @
z	100	100	100	100	100	100
m	100	100	100	100	80*#@	100
n	95	100	100	100	95	100
l	100 #	70* @	95 #	100 #	15*#@	30*#@
r	100	95	100	100	100	100
w	95	100	95	100	100	70*#@
j	100	95	100	100	100	100
i:	95	100	95	85	100	100
ɪ	100	100	100	90	100	80*#@
æ	100	95	100	100	100	100
a:	100	95	100	100	100	95
ʌ	100	95	95	95	95	40*#@
ɒ	95	100	100	100	100	95
o:	100	100	100	100	100	100
ʊ	100	95	100	100	95	85
u:	95	95	95	80	75*#@	80
ɜ:	100	100	100	100	75*#@	90
aɪ	95	100	100	90	100	100

Table 1. Percentage correct identifications for the various phonemes used in this study. Significant deviations from NATURAL, ZERO, and NATPHAS are indicated by *, #, and @ respectively.
(Underlined symbol: p<.01; Symbol without underline: p<.05)

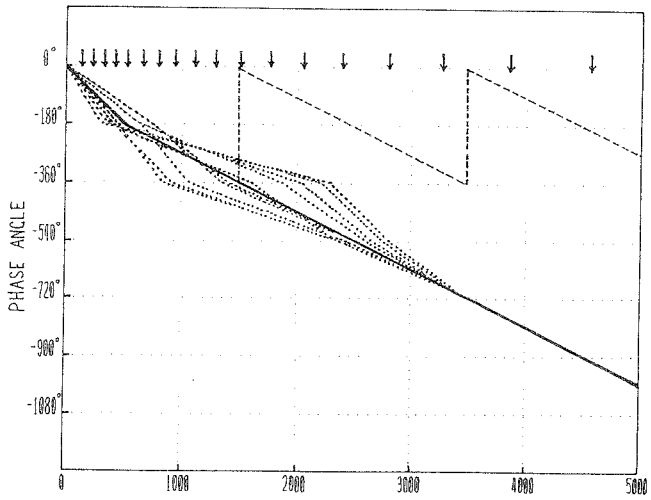

Figure 1. Solid line indicates the theoretical phase (after Fant, 1968) of a neutral vowel determined as -180° per formant. The dotted lines represent phase calculated in the same way for the 11 monophthongs of Australian English. The dashed line is the neutral vowel phase wrapped into the range 0° to -360° and the arrows indicate the points on the dashed line used to determine the phase of the filter in the 0° to -810° condition.

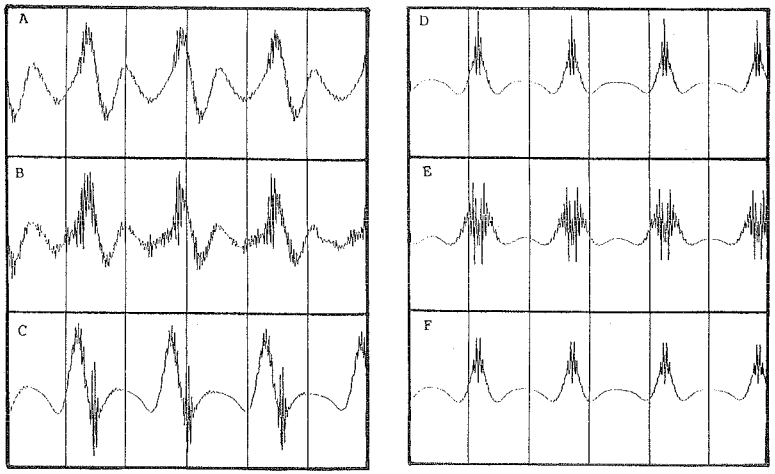

Figure 2. /i/

A: Natural speech; B: Vocoder amplitude plus Natural Phase;
C: Impulse delay; D: Zero phase; E: 0 to -810 degrees phase;
F: 0 to -360 degrees phase.