A COMPARISON OF THE INTELLIGIBILITY SCORES OF CONSONANTS AND VOWELS USING CHANNEL AND FORMANT VOCODED SPEECH

R.H. Mannell and J.E. Clark

Speech, Hearing and Language Research Centre
Macquarie University, Sydney, Australia

ABSTRACT

The intelligibility of various phonetic classes is examined following vocoding by a formant vocoder and various (Hz-scaled and Bark-scaled) implementations of a channel vocoder. The results suggest that particularly in the case of various consonant classes the 1 Bark channel vocoder performed a little (but significantly) better than the Hz-scaled channel vocoders and much better than the formant vocoder. The 1 Bark vocoder achieved intelligibity results equivalent to natural speech for most phonetic classes. The results support the idea that channel vocoding techniques are intrinsically capable of achieving natural speech intelligibity and suggest that formant systems may be intrinsically incapable of achieving natural intelligibity for certain phonetic classes.

1. INTRODUCTION

This work arises from a general interest in synthesiser performance and particularly in the performance of competing parametric encoding strategies. The limitations of speech synthesis performance is well recognised and there is a growing body of quantitative evidence [1,3,6,7] as to the nature of these performance limitations. Synthetic vowel intelligibility is often equivalent to that of natural vowels whilst consonants, on the other hand appear to consistently demonstrate a shortfall in synthetic relative to natural intelligibility [1,6,7]. Most of these studies examined the intelligibility of synthesis-by-rule and text-to-speech systems and all examined the performance of formant synthesisers. One question that this study attempts to address is whether these findings reflect merely our limited ability to formulate rules for the generation of consonants using a formant-based synthesis-by-rule system or whether there is a more fundamental limitation in the potential performance of formant synthesis with respect to consonants. An even greater motivation for the present study is an examination of whether channel synthesis also shares with formant synthesis any fundamental limitation in its ability to synthesise consonants and further, which filter configurations produce consonant intelligibility performance approaching that of natural speech. Vocoders were utilised in this study as they allow a direct examination of various encoding strategies without the confounding effects of rule and database defects potentially inherent in synthesis-by-rule systems. Further, vocoder software simulations are very flexible allowing easy modification of filter configurations etc.

The present paper is a further progress report on a study outlined at the Tallin conference [2].

2. PROCEDURE

The primary means for manipulating the parametric information content of the resynthesised speech was via a classical channel vocoder, first described by Dudley [4]. A channel vocoder was used because it is relatively free of any major a priori assumptions about the primacy or otherwise of particular spectral features such as energy peaks or depressions as bearers of phonological information. A channel vocoder comprehensively encodes all spectral components able to be resolved by the frequency resolution (bandwidth) of the filters. Being a vocoder the speech is resynthesised directly from spectral information extracted from a natural speech signal and not from rules and databases as would be the case with synthesis-by-rule systems.

The vocoder (figure 1) is a software simulation residing on a VAX 11/750 computer consisting of band pass (BP) and low pass (LP) FIR filters designed using the well known window synthesis technique. This allowed for considerable flexibility in filter design and numerous filters of different time and frequency domain characteristics have been designed and used over this project. The present paper will only examine a subset of these filter configurations in which the BP channel filter bandwidths are varied in both the Hz-scale and the Bark scale. The pitch and excitation detection algorithms were adapted to the limited input data and all decisions made by that module were confirmed by a experienced phonetician. It is unlikely that the pitch/excitation module contributed in any way to the final intelligibility results.

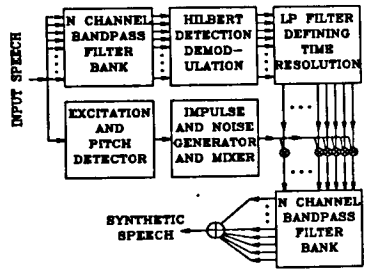

FIGURE 1. S.H.L.R.C. CHANNEL VOCODER

The speech was digitised and bandlimited to 0 to 5 kHz. Four Hz-scaled BP channel filterbank configurations (Bandwidths: 100, 200, 400 and 800 Hz) and five Bark-scaled filterbanks (Bandwidths: 0.75, 1.0, 1.5, 2.0 and 3.0 Bark) were utilised in the part of the project described in this paper. The outputs of these filters were passed through identical Hilbert transforms. These Hilbert filters were inserted before the following LP filters to maintain a constant spectrum envelope demodulator (deemed desirable for conditions in which the LP filter was varied to manipulate the time resolution of the total system). For the conditions reported upon in this paper the LP filters were fixed at 50 Hz (a time resolution of 10 msecs as defined by the sampling theory) and as this filter was the "slowest" filter in the system it defined the time resolution of all the systems as a constant 10 msecs. In all conditions the BP synthesis filters were identical to the BP analysis filters.

The formant vocoder used in this study was developed at the Joint Speech Research Unit (JSRU) in the United Kingdom and was based on an automatic formant analysis system described by Dupree [5] and coupled with the highly regarded JSRU formant synthesiser described by Rye and Holmes [8]. The time resolution of this system was the same as that used in the channel vocoder configurations (ie. 10 msecs).

The test items were 11 vowels in an /h_d/ frame and 19 consonants in a CV frame (V=/a:/) spoken by a native speaker of Australian English. These tokens were recorded to professional audio standards in an echo free room and digitised (16 bits) onto and vocoded on a VAX 11/750 computer. The tests were conducted in a sound treated room using calibrated TDH-49 headphones with standard cushions and circumaural seals. The level normalised test tokens were presented both in silence and in +6, 0 and -6 dB S/N (utilising USASI speech-shaped noise) at a presentation level of 70 dB s.p.l. (ref. 20 μPa). There were 20 different listeners for each condition all of whom were native speakers of Australian English, none of whom had any experience with synthetic speech, and none of whom had any history of hearing or speech pathology. All subjects were screened with a simple speech discrimination test which ensured that they were reliably able to identify monosyllabic words presented at 40 dB s.p.l. All relevant pairs of conditions were compared using the chi square test and tested for significant difference at the 0.01 level.

3. RESULTS AND DISCUSSION

The intelligibility results for the unmasked conditions are summarised in figures 2 to 5. Since the 0.75 and 1.0 Bark conditions are not significantly different for any phonetic classes the 1 Bark condition is the optimal Bark-scaled condition (ie. fewer filters for no loss of intelligibility) and is used in the following discussion as the Bark-scaled reference condition. The 100 Hz condition is significantly higher in intelligibility than the 200 Hz condition for some phonetic classes and so the 100 Hz condition is considered the optimal Hz-scaled condition.

The difference in vowel intelligibility between the natural condition and the 100 Hz and the 1 Bark conditions is not significant when presented unmasked but the performance of masked 100 Hz vowels is significantly lower than the performance of both natural and 1 Bark vowels. The formant vocoded vowels are moderately, but significantly, lower in intelligibility than the natural vowels. Further, the formant vocoded vowels are significantly lower in

intelligibility than the 100 Hz vowels which are in turn not significantly different from the 1 Bark vowels. It should be noted that intelligibility significantly deteriorates (for the vowels) when the frequency resolution drops below 200 Hz and 1.0 Bark.

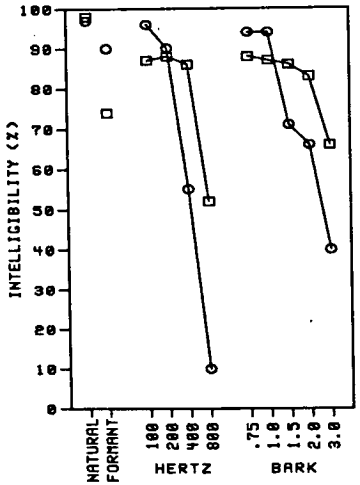

FIGURE 2. O ALL VOWELS
□ ALL CONSONANTS

The intelligibity of the formant vocoded consonants is considerably (and significantly) lower than that of the natural, 100-200 Hz and 0.75-1.0 Bark consonants. There is no significant difference between the 100 Hz and the 1 Bark unmasked consonant conditions. Consonant intelligibility does not deteriorate significantly up to 400 Hz and up to 2.0 Bark indicating that consonants are generally able to withstand poorer frequency resolution than are vowels.

In figure 3 it is clear that the intelligibity of channel vocoded affricates is unimpaired at all bandwidths whilst the formant vocoded affricates are significantly and markedly lower in intelligibility compared with all other conditions. The stops show a somewhat unpredictable pattern but there is little significant difference between the voiceless and voiced stops and so they will be dealt with here as a single class. Generally all vocoded stops are lower in intelligibility than the natural condition. The formant vocoded stops are generally lower in intelligibility than the channel vocoded stops, but not significantly lower than the 100 Hz and the 1 Bark conditions. Masked 100 Hz voiceless stops are significantly lower in intelligibity

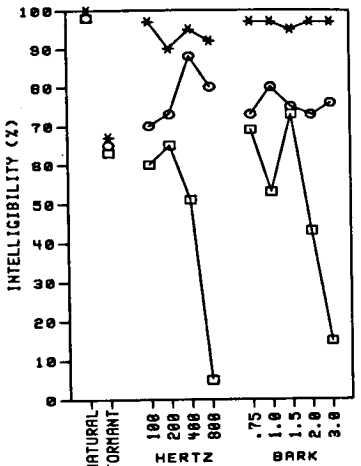

FIGURE 3. * AFFRICATES
O VOICELESS STOPS
□ VOICED STOPS

than masked 1 Bark voiceless stops. It is interesting to note that the voiceless stops do not deteriorate in intelligibity with increasing bandwidth whilst voiced stops do.

In figure 4 it can be seen that there is no significant difference in intelligibity for both voiceless and voiced fricatives between the natural condition and the 100-400 Hz and the 0.75-2.0 Bark conditions. The formant vocoded fricatives, on the other hand are significantly lower than the natural and the channel vocoded fricatives in intelligibity. There is no difference between the best Hz-scaled and Bark-scaled conditions for both unmasked and masked presentation. Only the voiced fricatives are effected by increasing bandwidth and only for the 800 Hz condition.

Figure 5 details the nasal and approximant results which display very similar patterns to each other and be treated in the following comments together. Firstly, (and predictably) the intelligibity curves behave very much like that of the vowels as bandwidth increases. The best Hz-scaled and Bark-scaled conditions are not significantly different from the natural intelligibity, and intelligibity drops off significantly when bandwidth exceeds 400 Hz or 2 Bark. The formant vocoded condition is not significantly less intelligible than natural or the best channel vocoded conditions when heard unmasked but is significantly lower in intelligibility than natural tokens when heard

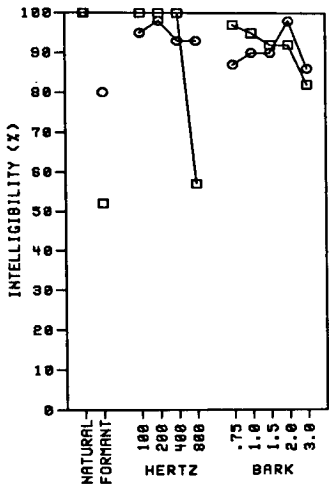

FIGURE 4. O VOICELESS FRICATIVES
□ VOICED FRICATIVES

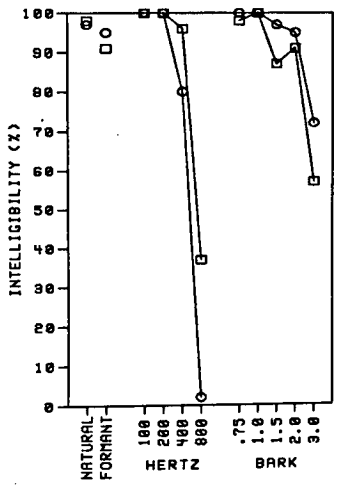

FIGURE 5. O NASALS
□ APPROXIMANTS

masked. The 1 Bark condition is not significantly less intelligible than natural for both nasals and approximants when presented either unmasked or masked. This is also true for the 100 Hz condition but with the exception that masked approximants are significantly lower in intelligibity than masked natural approximants.

4. CONCLUSIONS

This study presents evidence that channel vocoders with a 1 Bark bandwidth filterbank perform significantly better than formant vocoders. A 1 Bark filterbank vocoder is equivalent in intelligibity to natural speech for all phonetic classes except the stops and the approximants (and in the second case only when masked). It also performs marginally better than the 100 Hz filterbank vocoder (this is only evident when tokens are masked). It cannot be stated with complete confidence that formant systems are inherently less able to parametrically encode speech than 1 Bark channel vocoders (and channel vocoders in general) however this study supports that conclusion and there does not seem to be any evidence in the literature to support the opposite conclusion (particularly with respect to consonants).It seems likely that a 1 Bark channel vocoder has the intrinsic ability to adequately encode phonologically relevant parametric detail with sufficient accuracy to produce intelligibity approaching, or equal to, natural speech.

5. REFERENCES

[1] CLARK, J.E. (1983), "Intelligibility comparisons for two synthetic and one natural speech source", *J. Phon.* 11, 37-49.

[2] CLARK, J.E. MANNELL, R.H., & OSTRY, D. (1987), "Time and frequency resolution constraints on synthetic speech intelligibility", *Proc. XI ICPhS*, Tallin Estonia, Aug. 1-7, 1987.

[3] CLARK, J.E., DERMODY, P. & PALETHORPE, S. (1985), "Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons", *J.A.S.A.* 78, 458-462.

[4] DUDLEY, H. (1939), "Remaking speech", *J.A.S.A.* 11, 169-177.

[5] DUPREE, B.C. (1978), "Automatic formant analysis", *Proc. Institute of Acoustics*, Spring meeting, Cambridge.

[6] LOGAN, J.S., PISONI, D.B. & GREENE, B.G. (1985), "Measuring the segmental intelligibility of synthetic speech: results from eight text-to-speech systems", *Research on Speech Perception*, Progress Report No. 12, 319-334.

[7] PISONI, D.B., NUSBAUM, H.C., & GREENE, B.G. (1985), "Perception of synthetic speech generated by rule", *Proc. IEEE* 73, 1665-1675.

[8] RYE, J.M. & HOLMES, J.N. (1982), "A versatile software parallel-formant speech synthesiser", *JSRU Research Report* No 1016