SOME COMPARATIVE CHARACTERISTICS OF UNIFORM AND AUDITORILY SCALED CHANNEL SYNTHESIS

J.E. Clark and R.H. Mannell
Speech Hearing & Language Research Centre
Macquarie University

ABSTRACT - This paper examines the comparative phonetic level intelligibility characteristics of two channel vocoder type synthesis systems, one based on a uniform bandwidth filterbank, and the other on an auditorily scaled filterbank. The intelligibilty tests were conducted using listeners with no prior experience of synthesised speech, and employed masking noise to help expose differences in the perceptual robustness of the test corpus. The intelligibility of the natural input speech tested under the same conditions was used the benchmark for all comparisons. The results suggest that the Bark scale derived synthesis may have intelligibility characteristics closer to those of natural speech than the uniform filterbank synthesis, is perceptually more robust, and is more cost effective in its use of available channel encoding.

INTRODUCTION

Although techniques for the synthesis of speech are now well established and have begun to find some limited commercial application, all forms of parametric synthesis still exhibit limitations in intelligibility and quality relative to natural speech. The nature of these limitations as they relate to formant and l.p.c. coded speech are well documented in studies over more that 40 years, including those of Dudley et al (1939), Keeler et al (1976), Clark (1983), Pols & Olive(1983), Clark et al (1985), Pisoni et al (1985), and Malsheen et al (1987).

All these studies indicate that synthesised speech has difficulty in reaching the intelligibility levels of natural speech, particularly when listeners are presented with single word or syllable test tokens in open set response test formats. Moreover, there is also evidence in these studies that synthesis is less robust in noise than natural speech, and that it makes greater processing demands on the listener. The problem of quality is less extensively documented, but it is evident from current trends in the use of computer speech output in public domain information technology applications that stored natural speech is still often preferred to synthesis in many instances.

Explanations for these deficiencies in synthesis are argued by some researchers as primarily due to deficiencies in the parametric data controlling the synthesis model. As a result, much effort in recent years has been devoted to improving the quality of such parametric information, especially where it is generated via linguistic rules. This viewpoint assumes that the synthesis model itself is adequate, which if untrue to any appreciable degree, will firmly delimit the cost-benefit of attempts to improve parametric data, and consequent improvements in synthesis intelligibility and quality.

OBJECTIVES

The work reported in this paper is part of a larger examination of the encoding requirements of parametric synthesis. The basic question it seeks to explore, is just what degree of delicacy is required in the parametric description of the time-varying speech spectrum in order to generate synthetic speech output with an intelligibility and robustness which approaches that of natural speech. This question of parametric description adequacy also has relevance to speech processing strategies for other applications such as prostheses for the hearing impaired, and automatic speech recognition.

The specific objectives of the work reported here were to examine the relative intelligibility properties of speech synthesised using equal bandwidth and Bark scale derived filterbanks, and to compare both with the intelligibility of the natural speech input.

METHODOLOGY

The speech synthesis was generated using a conventional channel vocoder system of the kind described by Schroeder (1966) and whose principles of operation are well known. The vocoder was implemented in FORTRAN on a VAX 11/750 and used identical conventional FIR bandpass filter structures for the both the analysis and synthesis filterbanks. The sampling rate was 10KHz, and the resynthesis frame rate was 10mS. A channel vocoder was chosen for this purpose because it allows the time-varying energy of the speech spectrum to be described in a comprehensive fashion which makes no assumptions about what might or might not be phonologically salient features of that spectrum. The block diagram of the vocoder is shown in Fig. 1.

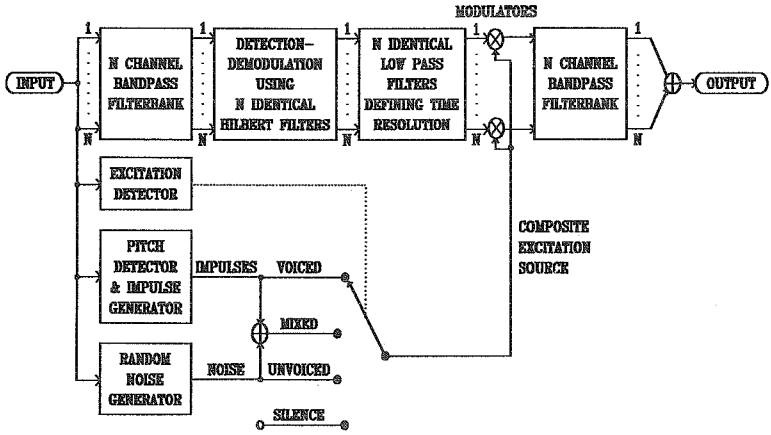

FIGURE 1

By varying the number of filters in the filterbank while maintaining a constant overall frequency range of 4800Hz, it was possible to alter the detail (or resolution) with which the time-varying spectrum was specified and resynthesised. In the uniform bandwidth case, bandpass filters of 100, 200, 400, and 800Hz were used, yielding 48, 24, 12 and 6 channel vocoder systems respectively. In the auditorily scaled case, the bandwidths of filters were derived from the bark scale, which relates acoustical frequency to perceptual frequency resolution of the human auditory system (Zwicker & Terhardt, 1980). The effective result is a rapid widening of filter bandwidth above the region of 1KHz.

For a given number of channels in the vocoder, the parametric description of the time-varying energy in the speech spectrum was thus either linearly or psychophysically distributed over its 0 - 4800Hz range. In the case of the auditorily scaled filterbanks, increase in the bandwidth of the individual filters of the filterbank with increasing frequency results in increasing spectral distortion which gives greater amplitude to high frequency components of the signal in resynthesis than were present in the input. When generating the test stimuli described below, two complete ensembles were produced. One set ignored the presence of the distortion effect, and the other included a simple weighting correction in the analysis output to account for the high frequency pre-emphasis caused by spectral distortion, such that a spectrally flat input signal produced a spectrally flat synthesis output.

The original natural speech tokens used as input to the synthesis system (and as benchmark intelligibility data) in generating the various test stimuli, consisted of a set of 11 vowels in /h-d/ frames, and a set of 19 consonants in CV syllables. They were recorded by a male speaker in a sound treated room to professional audio standards. The use of nonsense syllables was dictated by the need to ensure that the listeners made minimal use of linguistic context in identifying the test tokens.

For each filterbank type, and the original input speech, four randomised sets of test stimuli with different masking conditions were prepared. These were; quiet (no masking), +6dB, 0dB and -6dB signal to noise ratio masking. The masker employed was noise spectrally shaped to approximate the long term spectrum of male speech. Of the possible spectral shapes that might be chosen for the masker, speech shaped noise was considered the most appropriate in maintaining a relatively consistent masking effect over the whole speech spectrum. The masker and test stimuli intensity ratios were computed individually for each token using a speech editing utility, and the complete file normalised to a predetermined signal level.

The listeners used for the intelligibility trials were all screened to ensure that they were able to identify similar speech tokens at a presentation level at least 30dB s.p.l. below that of the actual test materials. None had any known history of hearing disorder, and all were adult native speakers of Australian English. Separate groups of 20 listeners each were used to test each filterbank type to help minimise confounding effects caused by practice. The test tokens were heard in random order in /h-d/ and CV groups, in which the signal to noise ratio commenced with the quiet and ended with the -6dB condition. The tests were conducted in a sound treated room using TDH49 headphones with supra-aural cushions and circumaural seals at a presentation level of 70dB s.p.l. (ref. 20 uPa).

RESULTS AND DISCUSSION

The results for the uniform bandwidth filterbanks (see also Clark et al, 1987) are shown in fig. 2. It can be seen that the vowels show a rapid reduction in intelligibility below 24 channels. This reflects the significant demand made by vowels on frequency resolution in order to preserve the spectral peak information which has a substantial role in establishing their auditory identity. Although the 48 channel filterbank condition approaches equivalent intelligibility to natural speech, it is substantially more vulnerable to heavy masking.

The best consonant intelligibility levels are poorer overall than those for the vowels, with very little difference existing between the 24 and 48 channel filterbanks as already seen for the vowels. There is, however, nothing like the same reduction in intelligibility as the number of channels in the filterbank is reduced. Even the 12 channel filterbank has quite high intelligibility levels, although it is more strongly affected by noise masking than the 24 and 48 channel cases.

The unweighted auditorily scaled filterbank results are shown in fig. 3. In this form of filterbank, reduction of the number channels has rather less impact on vowel intelligibility overall, and also shows reduced vulnerability to masking. Both the 24 and 18 channel (0.75 and 1.0 bark respectively) versions show vowel intelligibility which approaches that of natural speech for all listening conditions. The consonants have intelligibility characteristics which appear slightly better than the best uniform filterbank results, and are less vulnerable to masking. The 6 channel version is, however, substantially better than its uniform bandwidth counterpart.

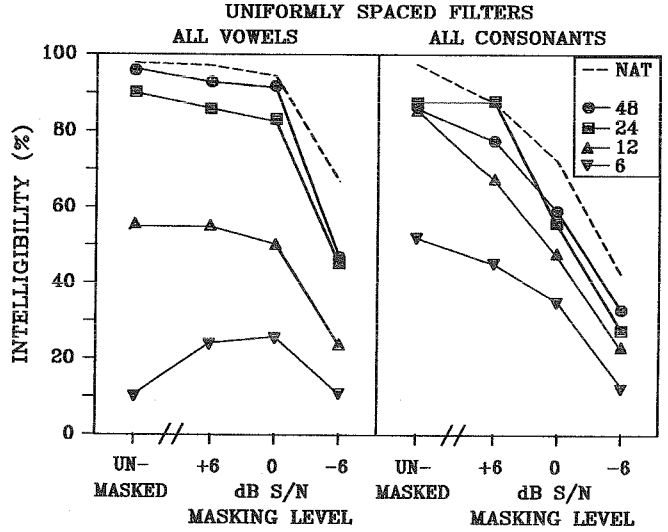

FIGURE 2

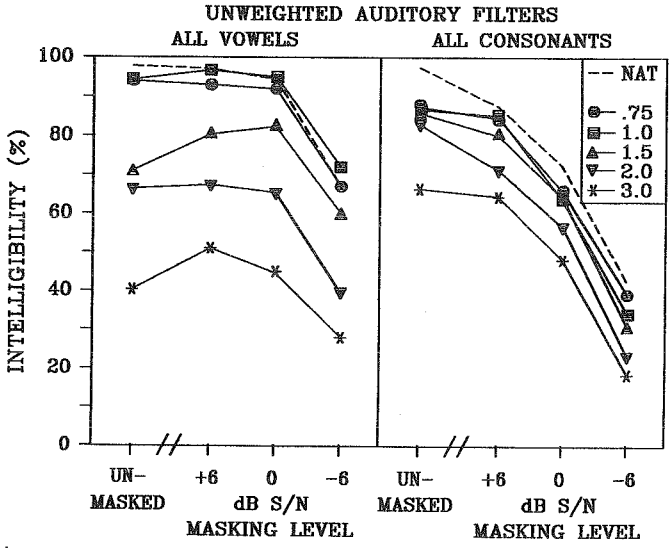

FIGURE 3

The weighted auditory filterbank results are shown in figure 4. The intelligibility scores for the vowels in quiet listening conditions are overall slightly higher than for the unweighted version, although there is little difference between versions for the consonants. However the weighted version scores show its strikingly greater vulnerability to masking than the unweighted counterparts for all noise masked listening conditions in both vowels and consonants. The explanation for this appears to lie in the fact the weighting procedure reduces the height of spectral peaks as well as reducing overall spectral tilt. The result is synthesis which has improved naturalness, but a spectral structure whose important information bearing aspects are more strongly affected by a given signal to noise ratio than is the unweighted version.

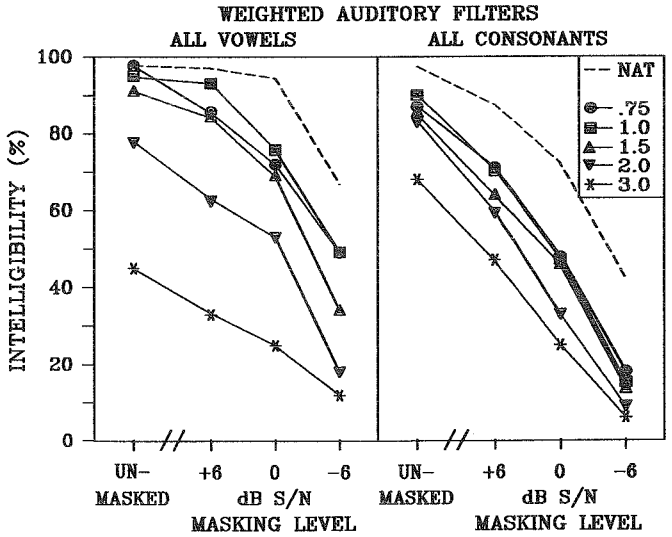

FIGURE 4

CONCLUSIONS

As might be expected, there seem to be some advantages in representing and resynthesising the speech signal with an auditorily scaled parametric description of its time varying spectrum. These advantages include more consistent intelligibility levels between vowels and consonants and more effective encoding of phonological information, especially in 6 to 12 channel systems. The unweighted auditorily scaled filterbank version showed good robustness in noise, but the high quiet listening scores of the weighted version suggest that the investigation of a more effective weighting procedure might produce worthwhile gains in naturalness without sacrificing robustness. It is clear from the results presented here that auditorily scaled filterbank encoding and resynthesis using 12 or 18 channels can approach natural speech intelligibility, and even 6 or 9 channels will encode useful information where a much simplified signal representation might be needed, as in as hearing prostheses such as tactile devices or cochlear implants.

REFERENCES

Clark, J.E. (1983) "Intelligibility comparisons for two synthetic and one natural speech source", *J. Phonetics*, 11, 37-50

Clark, J.E., Dermody, P. & Palethorpe, S. (1985) "Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons", *J. Acoust. Soc. Am.*, 78, 458-462

Clark, J.E., Mannell, R.H. & Ostry, D. (1987) "Time and frequency constraints on synthetic speech intelligibility", *Proc. XIth ISPhS*, 28.2.1, 215-218

Dudley, H. et al (1939) "A synthetic speaker", J. Franklin Inst., 227, 762-763

Pisoni, D.B., Nusbaum, H.C & Greene, B.G. (1985) "Perception of synthetic speech by rule", *Proc. IEEE*, 73, 1665-1676

Keeler, L.O. et al (1976), "Two preliniary studies of the intelligibiltiy of predictor-coefficient and formant coded speech", *IEEE Trans. ASSP*, ASSP-24, 429-432

Malsheen, B.J et al (1987) "Intelligibility of English, French, German, and Spanish consonants generated by rule over simulated telephone bandwidths", *Proc. XIth ISPhS*, 28.1.4, 211-214

Pols, L. C.W. & Olive, J.P. (1983) "Intelligibility of consonants in CVC uttertances produced by dyadic rule synthesis", *Speech Communication*, 2, 3-13

Schroeder, M.R. (1966) "Vocoders, analysis and synthesis of speech", *Proc. IEEE*, 54, 720-733

Zwicker, E. & Terhardt, E. (1980) "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", *J. Acoust. Soc. Am.*, 75, 219-223