TIME AND FREQUENCY RESOLUTION CONSTRAINTS ON SYNTHETIC SPEECH INTELLIGIBILITY

J.E. CLARK

Speech Hearing & Language
Research Centre
Macquarie University
North Ryde, N.S.W. 2113
Australia

R.H. MANNELL

Speech Hearing & Language
Research Centre
Macquarie University
North Ryde, N.S.W. 2113
Australia

D. OSTRY

Radiophysics Division
C.S.I.R.O.
North Ryde
N.S.W.
Australia

ABSTRACT

The effects of time and frequency resolution properties of resynthesised natural speech on its intelligibility were investigated at the phonological level. An automatic analysis-resynthesis channel vocoder was used to manipulate the time and frequency properties of the synthetic speech. The original natural speech and a high quality formant vocoder provided the comparative performance benchmarks. The test materials were noise-masked monosyllables. Results showed that vowels made the greatest demands on frequency resolution, with both consonants and vowels showing similar overall demands on time resolution. The higher information rate channel vocoders were markedly superior in consonant intelligibility to the formant vocoder benchmark.

INTRODUCTION

This investigation was motivated by a general interest in the performance of speech synthesis systems, and in the parametric coding required to represent the phonologically related information content with perceptual adequacy.

Limitations in the intelligibility and perceptual robustness of synthesised speech have been observed since the time of Stewart [1]. There has been accumulating quantitative evidence of this limitation in more recent times [2], [3], [4], [5], prompting Pisoni et al [4] to comment that "..it seems more advisable to use a low-cost synthesizer to provide spoken confirmation of database entries than as a voice response system in the cockpit of a jet fighter or a helicopter." (p.1675).

OBJECTIVES

The broad objectives of this study were:

1. To try and determine some of the ways in which the intelligibility of synthesised speech is constrained by resolution of the information (in its time and frequency domains) of the information contained in its resynthesis parameters.

2. To relate the findings on synthesis parameter manipulation to the intelligibility of the original natural speech and a known high quality formant vocoder as benchmark comparisons.

METHODOLOGY

Speech Processing Systems

A classical channel vocoder was chosen as the means for manipulating the parametric information content of the resynthesised speech signal. This class of vocoder has time and frequency resolution properties which are explicit in their structure. Moreover, they make few apriori assumptions in their parametric encoding about the nature of the phonologically related information bearing properties of the time-varying spectrum of speech signals. They do, of course, make some necessary assumptions in relation vocal tract excitation sources, about the nature of its periodicity and aperiodicity.

The channel vocoder is the earliest electronic speech analysis-resynthesis device. It was first developed some 50 years ago, motivated by an interest in reducing telephone transmission bandwidths. This is achieved (without great coding efficiency) by only transmitting the relatively slow time-varying changes in the energy envelopes of the speech signal spectrum as sampled by a filter-bank analyser spanning the range of frequencies of interest in the signal to be processed. The output of each analysis filter is detected and processed to produce the necessary slow time-varying envelope signal, and this information is then transmitted for resynthesis at the other end of the transmission path. The resynthesis is achieved using a corresponding filter-bank excited by periodic and/or aperiodic functions of uniform spectral energy, or a mix of both, as appropriate The actual excitation level appropriate to each filter is set using a multiplier controlled by the energy envelope signal derived from the corresponding analysis filter channel. Excitation function information defining whether it is periodic or not, and in the former case the period itself, is derived direcly from the input signal and transmitted to the resynthesis component of the system separately. Fig. 1 shows the structure of the vocoder in schematic form.

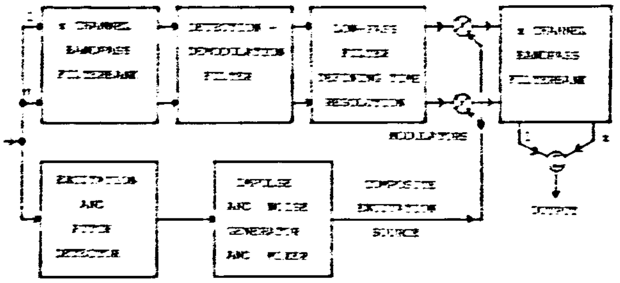
n CHANNEL BANDPASS FILTERBANK
DETECTION - DEMODULATION FILTER
LOW-PASS FILTER DEFINING TIME RESOLUTION
MODULATORS
n CHANNEL BANDPASS FILTERBANK
INPUT
EXCITATION AND PITCH DETECTOR
IMPULSE AND NOISE GENERATOR AND MIXER
COMPOSITE EXCITATION SOURCE
OUTPUT

Fig.1 S.H.L.R.C. Research Vocoder

The vocoder was realized as software on a VAX 11/750, and makes no attempt to meet any particular criteria of encoding or computational efficiency, given that it is only intended as a signal manipulation device. Identical filterbanks were used for analysis and resynthesis.

Despite the venerable age of this speech processing device, there are several quite basic questions about its design parameters which are not clearly resolved in the literature. It is not the purpose of this paper to discuss vocoder design, but it is worth noting that in develping the vocoder used in this present investigation, several different analysis-resynthesis filter types were tried together with several forms of analysis filter energy detection before settling on the configuration used in this investigation.

Opinions in the literature [7], [8], on requirements for analysis/synthesis filter properties vary. Despite some claims that it is desirable to use filters with relatively shallow skirt slopes and having well damped impulse responses, and that filter skirt response overlap is relatively unimportant because of the large amount of correlated energy occurring in adjacent bands, it was found in this study that such filters produced speech of unacceptable quality and intelligibility. By contrast, each of several filters tried with steep skirt slopes and much more restricted response overlap produced far better speech quality.

The effective frequency resolution of the system is set by the number of filters, and may be selected from 6, 12, 24, and 48, to give uniform bandwidths of approximately 800, 400, 200 and 100 Hz respectively. The sampling rate of the vocoder is 10KHz, and hence the frequency range of the filterbank is 0 - 5KHz in all cases.

The question of optimal criteria for filterbank energy detection systems also seems unresolved in the literature. For the present investigation, the need for independent manipulation of the vocoder data rate made a Hilbert filter a suitable choice to meet the output ripple and response speed criteria. This is rather specific to the uses of the vocoder, being designed to provide an output energy envelope with a maximally rapid impulse response.

A separate set of low pass filters were used to limit the bandwidth of the energy envelope signals, so simulating changes in the vocoder data transmission rate. This controls the effective time resolution of the information transmitted for resynthesis. The cut-off of this filter may be set to give effective parameter update rates of 10, 20, 40, and 60mS. It may also be bypassed to give a limiting vocoder time resolution set by the combined effects of the analysis filterbank and energy detection systems.

Algorithms for deriving pitch and voicing status excitation data abound; the scheme used here is not claimed to have any special merit, but was a time-domain type specifically tailored to the needs of this vocoder. The excitation signals for resynthesis in the vocoder are derived by direct extraction of smoothed pitch data, and a voicing detection system which determines whether the signal is periodic, aperiodic, or a mixture of both. The detection system contains hysteresis to minimise voicing decision jitter.

The formant vocoder used was a standard high quality system at the Joint Speech Research Unit, using a copy of the master recording of the benchmark natural speech materials as input. The resynthesis uses a four formant systems based on the well known J.S.R.U. synthesiser, chosen because of its reputation for very high quality speech output.

Listening Tests

The perceptual properties of the acoustic speech signal were tested in conditions of near minimal linguistic context to minimise the confounding effects of top-down processing by listeners. A set of 11 /h-d/ words and 19 CV nonsense syllables representing a selection of the common vowels and consonants of English respectively, were employed.

Natural speech versions of the test materials that were used as input to the vocoder were tested to provide a benchmark for the vocoder speech intelligibility data. The original natural speech was recorded to professional standards in an echo free sound treated room. The full range of time resolutions were tested using a 24 channel vocoder, and the full range of frequency resolutions were tested using a 10mSec time resolution (data update rate). The formant vocoded speech was processed with a 10mSec time resolution.

All the speech types were tested unmasked, and masked, the latter at signal to noise ratios of +6, 0 & -6 dB. The masking noise had a sloping spectrum approximating the long term spectrum of male speakers of English, and all the test stimuli were level normalised using the standard Leq method. The test stimuli presentations were all randomised and recorded with a 500Hz tone preceding each stimulus, and an interstimulus interval of 5 Sec. The stimulus and test tape generation was done digitally on the VAX 11/750. Listeners were drawn from amongst students and staff at Macquarie University. No listeners experienced with the task or with speech synthesis were employed, and listeners were not used for more than a single test session. Prior to the test sessions listeners were given a simple speech discrimination test to ensure that they could accurately identify common monosyllabic words down to a presentation level of 45dB s.p.l. before being included in the test crew. All the test materials were presented at +70dB s.p.l. using TDH49 headphones with standard cushions and circumaural seals in a sound treated room.

Analysis Procedures

The response data was entered into a computer program which produced intelligibility scores by individual test condition, and pooled intelligibility scores.

RESULTS

Fig. 2 shows intelligibility by frequency resolution by masking. The vowels are overall more resistant to masking than the consonants, with the formant vocoder vowels the most resistant of all. Both the 48 and 24 channel vocoders produce highly intelligible vowels at all but the deepest masking condition, whilst the poor performance of the 12 and 6 channel vocoders demonstrates the importance of frequency resolution. Note the rising intelligibility with noise in the 6 channel case.

The synthesised consonants show lower overall intelligibility than the vowels, although the 48 and 24 channel vocoders show a resistance to masking which is comparable to or better than natural speech in conditions of moderate masking. The formant vocoder is a little poorer than the 12 channel vocoder, except in moderate to heavy masking.

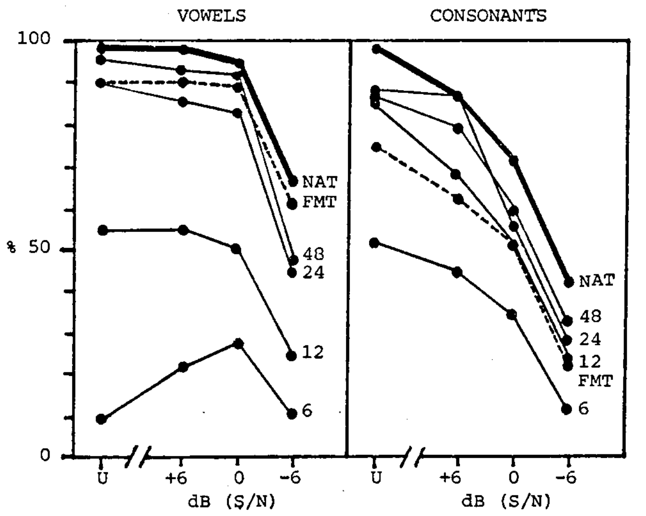

Fig.2 Intelligibilty X Frequency X Masking

Fig. 3 shows intelligibility by time resolution by masking. The vowels are relatively tolerant of reduced time resolution with no degradation until the 40mS condition, and a slight rise in intelligibility with moderate masking. The consonants show a similar pattern but with more rapid degradation at 40 and 60mS. The 10mS condition is least resistant to moderate masking. The formant vocoder has a performance which is comparable to or slightly better than a 40mS 24 channel vocoder.

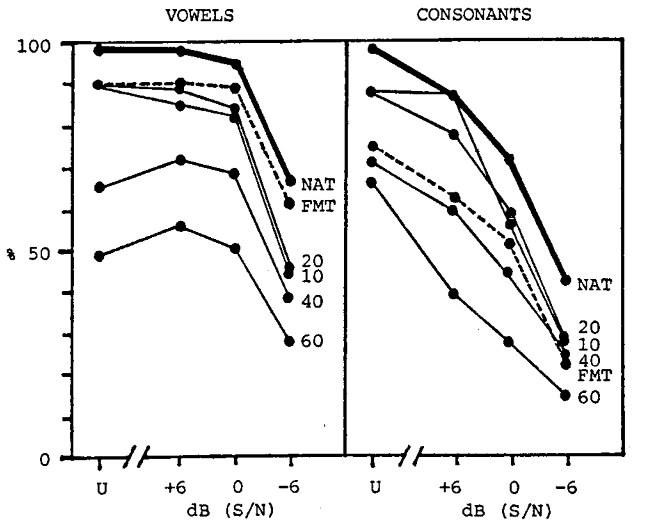

Fig.3 Intelligibility X Time X Masking

Figs. 4 & 5 show intelligibility by frequency and time resolution respectively, with pooled masking data. Vocoder vowel intelligibility decreases rapidly below 24 channel frequency resolution, and requires very high frequency resolution to approach formant vocoder performance. Consonant performance is far more tolerant of reduced frequency resolution, and suggests the formant vocoder to have a performance similar to that of a 12 channel vocoder.

Time resolution effects on performance are more consistent for both vowels and consonants, with appreciable reductions in intelligibility occurring at 40mS and above.

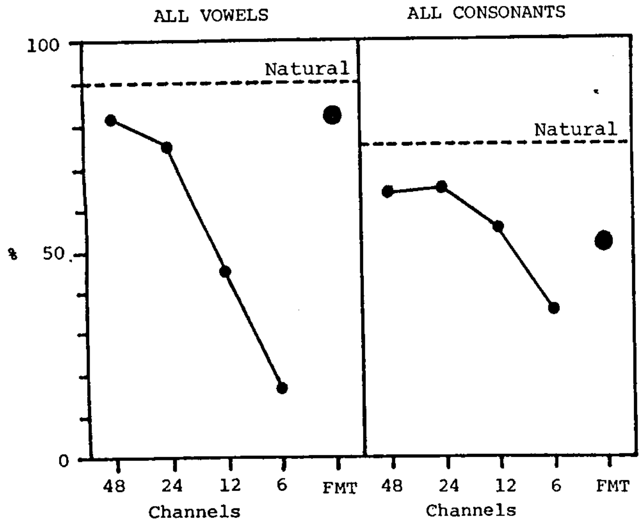

Fig.4 Pooled Intelligibility X Frequency

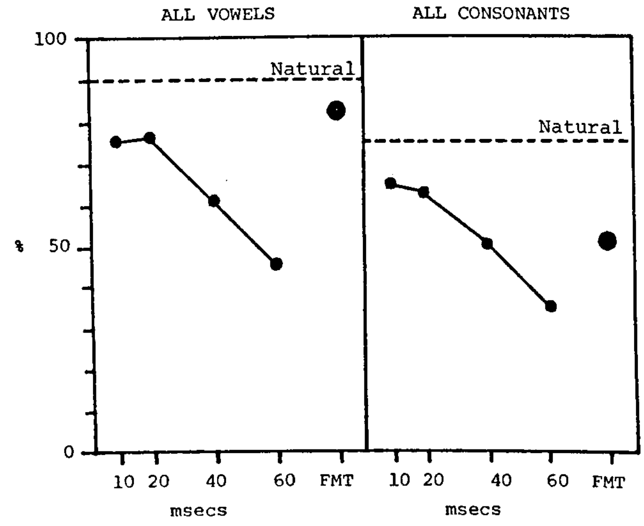

Fig.5 Pooled Intelligibility X Time

CONCLUSIONS

1. Overall intelligibility is more degraded by reduction in frequency resolution than by reduction in time resolution under the conditions tested (insofar as the two domains can be compared).

2. The comparative intelligibilities of vowels and consonants are reversed by progressive reduction in frequency resolution, but not time resolution. This illustrates the more stringent demand on frequency resolution in vowel parameter coding.

3. Time resolution reduction has a more consistent effect overall than frequency resolution reduction.

4. The formant vocoder shows a much greater performance differential between vowels and consonants than the channel vocoders. This generally poorer performance of consonant intelligibility with the formant coded speech suggests that it has appreciably less adequate parametric coding of consonantal information content than the 48 and 24 channel vocoders.

ACKNOWLEDGEMENTS

This work has been supported under the Macquarie-C.S.I.R.O. research grant scheme, and the A.R.G.S. The contribution of the technical staff of the S.H.L.R.C. is gratefully acknowledged, and we would like to thank Alison Brierley for running many of the listening test sessions, and tabulating the listener response data.

REFERENCES

[1] J.Q. Stewart, "An electrical analogue of the vocal organs", Nature, 110, 311-312, 1922.

[2] R. Carlson et al, "Evaluation of a text-to-speech system as a reading machine for the blind", STL-QPSR 2-3, 9-13 1976.

[3] J.E. Clark, "Intelligibility comparisons for two synthetic and one natural speech source", J. Phonetics, 11, 37-49 1983.

[4] D.B. Pisoni et al, "Perception of synthetic speech generated by rule", Proc. IEEE, 73, 1665-1675, 1985.

[5] J.E. Clark et al, "Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons", JASA, 78, 458-462, 1985.

[6] H. Dudley, "Synthetic Speech", Bell Labs Record, 15, 98-102, 1936.

[7] M.R. Schroeder, "Vocoers: Analysis synthesis of speech", Proc. IEEE, 54, 720-734, 1966.

[8] I.H. Witten, Computer Speech, Academic Press, 1982.