Proceedings of The Institute of Acoustics

THE EFFECT OF LIGHTWEIGHT PARTITIONS ON THE TRANSMISSION OF SYNTHETIC SPEECH

M. WEST

UNIVERSITY OF SALFORD

1. Introduction

Lightweight partitions are used extensively in schools, offices and dwellings. There are an enormous number of such partitions commercially available but little appropriate acoustic data. Invariably the partitions separate areas containing sources of speech and have often been selected using the Articulation Index (AI) Method (1). In many of these cases the occupants of the rooms on either side of the partitions express levels of dissatisfaction so great as to suggest that the original design criteria may be totally incorrect. Before condemning the AI method outright two important factors which contribute to the above dissatisfaction should be considered. The first is flanking transmission which can seriously degrade partition performance. In this study only in situ panel transfer characteristics are used to avoid the problem. The second difficulty is the 'emotional content' in speech, which is of course impossible to include in any design index.

Even with these factors considered the AI method is unsuitable for the prediction of speech privacy and it is the object of this project to lay the foundations for a new method more closely related to the panel's effect on aspects of the acoustic signal more pertinent to speech perceptibility. The new technique is based on the corruption of formant transitional information in synthetic C-V segments of speech. Background noise effects have for the moment been ignored.

2. What is Wrong with the AI Method

AI was developed for use in telephony (2) and was originally based on a signal to noise measurement in each of 20 'critical' frequency bands.

$$S/N \text{ (dB peak)} = \text{Source level (dBrms long)} - \text{Attenuation} - BC(\text{dBrms long}) + 12$$

The source level is based on an rms averaged over a long period. The value used in each band is taken from composite male and female speech data. The 12dB is added to give a peak S/N and arises from the difference between the above source level and the L1 (average peak level exceeded for 1% of the time based on the average of many $\frac{1}{8}$th second L1's obtained during uttered syllables avoiding silences). The signal to noise values obtained in each frequency band are multiplied by weighting factors which are supposed to allow for the relative importance of each band to the perceptibility. The weighted band articulation values are then added to give the total articulation index. Numerous other weighting factor sets have been developed based on more practical frequency bands and the method has been applied to privacy estimation (3).

The use of the averaged S/N removes the possibility of inclusion of any of the time dependent features of speech known to have a significant effect on perceptibility. Moreover the use of a S/N presupposes that signal amplitude in a given band is related to perceptibility. The evidence is to the contrary (4) and suggests that the relative amplitudes of frequency components during an utterance are not important but rather it is the changes in frequency content which carry the perceptible information. These changes though similar in gross terms in male and female speech differ considerably in terms of fine spectral structure which is also perceptively important.

Although the above weighting factors have been obtained from percentage syllable articulation with low and high pass filters, they must be applied in isolation to each pass band, the effect on adjacent bands being assumed included in the original weightings. The method therefore constrains our view of a given speech signal to fixed bands and does not allow an evaluation of the effect of the attenuating path on formant structures.

3. Measurement of Steady State Panel Transfer Characteristic

Two large loudspeakers driven by a white noise generator are placed in the corners of the room opposite the partition. Semi reverberant rooms were chosen and no attempts were made to correct for reverberation time. Tape recordings were made at five locations in each room and these were subsequently processed to give an average spectral difference between the rooms. This is used to set up a precision spectrum shaper into which the synthetic utterances are fed.

The steady state method has been chosen for its simplicity and repeatability. Many workers in this field have used impulsive techniques for measurement of the transfer characteristic. There is however no specific advantage in using this technique for speech signals as is often claimed. For this particular programme the steady state measurement may even be preferable since we are concerned with formant transitions in which the time variation of the poles is sufficiently slow (between 20 & 400ms) for the panel to give its steady state response.

In general the transfer characteristics of lightweight partitions are fairly smooth functions of frequency and although a narrow band format is routinely used for the input and output signals to reveal the perceptively important formant features, a 1/3rd octave spectrum shaper adequately represents most panels.

4. Production of Synthetic Utterances

Synthetic speech has been chosen in preference to the natural speech normally used (a) because of its absolute repeatability (b) because it permits a much more accurate evaluation of the effect of the panel on the chosen perceptive feature (viz formant transitions.) There is a great deal of evidence that formant transitions play a key role in speech perception (4, 5) so much so that early speech synthesisers were based entirely on

them.

An OVE IIId serial speech synthesiser was used since the formants are of the simplest type with preset bandwidth and relative amplitude. A fixed speech fundamental of 100Hz was used to remove the contribution of intonation. Speech stress is removed by maintaining a level vowel amplitude although amplitude transitions between consonant and vowel are permitted. These factors have an important role in speech perception and will form the basis of the next phase of this investigation.

A precision computer based analysis system has been developed to extract the formant data from the natural speech C-V's. (5). This is then converted to a form suitable for use with the OVE which is also driven by the same computer system. In effect we have a vocoder. The resulting synthetic speech is of course much less perceptible than the natural speech from which it is derived.

The stimuli have been restricted to C-V's for the first set of experiments primarily to avoid complications of familiarity found with short words or logatoms.

5. Perceptibility Testing

After a preliminary set of tests using different methods the same/different discriminability judgement was found to be most useful. ("Are these stimuli the same or different").

The stimuli presented are alternately those passed through the partition and the original. These were presented in random order to 10 subjects without any previous phonetic training. It was also intended that decision latency be measured and included in the evaluation of perceptibility. This however turned out to be variable from subject to subject and did not appear to be meaningfully related to the perceptibility. The stimuli are presented at a rate of one pair every 2 seconds allowing only sufficient time for a single decision.

References

1. American National Standard ANSI S3.5 (1969)

2. French, N.R. and Steinberg, J.C. Factors Governing the Intelligibility of Speech Sounds. J. Acoust Soc Amer. 19, 1, (1947), 90-119

3. Youngs, R.W Revision of the Speech Privacy Calculation J. Acoust Soc Amer 38(4), (1965) 524-530

4. Cooper, F.S et al. Some Experiments on the Perception of Synthetic Speech Sounds. J. Acoust. Soc Amer 24 (1952) 6, 597-605

5. Kewley-Port D. Measurement of Formant Transitions in Naturally Produced Stop Consonant - Vowel Syllables. J. Acoust Soc Amer. 72(2),(1982), 379-389