Generating artificial noise for testing speech recognition systems

Problem presented by

Dave Fish and Simon Roberts

Jomega Ltd, Austrey, Warwickshire

Problem statement

The Study Group was presented with the problem of determining how to generate noise samples appropriate for reliably testing the quality of in-car speech recognition systems. Speech-driven applications are becoming more common in cars and it is important to be able to test how well such systems perform and to identify those aspects of the noise which may cause difficulties for the systems. This requires the characterisation of the noise environment within a car. These characterisations must then be used to efficiently compress the information so that testing can be performed economically.

Study Group contributors

David Allwright (Smith Institute)
Melvin Brown (Smith Institute)
Richard Eyres (University of Bristol)
Tom Griffin (University of Bristol)
Sam Howison (University of Oxford)
Ralf Jacobs (University of Strathclyde)
Colin Please (University of Southampton)
Richard Rosing (University of Lancaster)
Eddie Wilson (University of Bristol)

Report prepared by

David Allwright (Smith Institute)
Melvin Brown (Smith Institute)

1 Introduction

The Study Group was presented with the problem of determining how to generate noise samples appropriate for reliably testing the quality of in-car speech recognition systems. Speech-driven applications are becoming more common in cars and it is important to be able to test how well such systems perform and to identify those aspects of the noise which may cause difficulties for the systems. This requires the characterisation of the noise environment within a car. These characterisations must then be used to efficiently compress the information so that testing can be performed economically.

2 Basic categories of noise

The noise was classified into three main categories: "Broad-band", "Narrow-band" and "Transients". The first two categories were distinguished as follows:

The broad-band noise consists of random noise and is in general representative of the higher frequencies found within a car. The behaviour can in general be identified by the spectrum that is generated with a slope of the log(amplitude) to log(frequency) plot being around -1.3. Given this decay behaviour with frequency this source of noise would then be closely correlated to measures such as RMS, loudness or intelligibility.

The narrow-band noise comes from the various structural resonances within the car and from the forcing terms due to engine and gearbox. The resulting behaviour can in general be characterised again by a spectrum but in this case the position of peaks and the amplitude of the peaks are the dominant characteristics that determine the nature of the noise. The narrow-band frequencies of the structural resonances are essentially fixed, while those due to the engine and gearbox are speed-related. Others, such as the resonant frequencies of the air in the tyres, are temperature-related.

Transient noise covers noise due to, for instance, stone impacts, and driving over a catseye, road marking or pothole; and also internal transient noises such as direction indicator noise and windscreen wiper transients (on reversing direction at the extremes of the sweep).

3 Methods of synthesis

3.1 Broad and narrow band noise

The idea of using randomised phase Fourier synthesis fitted over different bands of the frequency spectrum to capture both broad-band and narrow-band features was considered. Monomial frequency dependence of signal magnitude was initially assumed with a parameterised power-law. Such parameters might characterise different sources. There was considerable discussion on how to avoid any "beat" behaviour within the reconstructed noise, and avoid the periodicity resulting from the discrete Fourier transform. If the reconstructed signal has the form

$$\sum_i A_i \sin(w_i t + \phi_i)$$

with amplitude A_i, frequency w_i and phase ϕ_i, then the following suggestions were considered reasonable methods:

1. Taking $w_i = i\omega$, where ω is a fundamental frequency, the phase could be altered in a chaotic manner on a timescale longer than the fundamental frequency. For example ϕ_i could be taken to be time-dependent and be determined by a chaotic oscillator (*e.g.* the Rössler Oscillator in a suitable parameter regime).

2. The frequencies chosen could be nearly harmonic with

$$w_i = i\omega + c_i \,,$$

   where the c_i are random numbers chosen before the simulation begins.

3. The w_i could be chosen randomly from across the relevant frequency spectrum and the A_i could then be modified to ensure the necessary amplitude or power spectrum distribution is achieved.

In addition to these ideas on reconstruction, topics of envelope modulation and parameter sensitivity were also explored. In addition to the dependencies of the envelope on speed, there are external factors that can raise the envelope such as driving under a bridge or by a wall.

3.2 Transient noise

The final category was transient noise and here there was a less clear view on how such transients should be characterised. The idea of defining a basis space for fitting vehicle noise was discussed along with the need to define measures of sufficiency in defining sample lengths for signals. Such basis spaces and measures were conjectured to depend on the ultimate use of the reconstructed signal. There was considerable comment that the basis space that was adopted should not be "close" to basis states used by many current voice recognition systems since this might taint the results of experiments being considered. The idea of using this space to define the noise environment domains of acceptable operation of different speech analysis systems for a given source of speech was explored. This could be used as a comparative method of assessing different systems.

4 Delay space methods

There was discussion of the delay space method for capturing nonlinear signals. Issues of sample period, embedding dimension and sample size arose. Moreover further advice was required on how to synthesise signals from such models. Such spaces could also be used to look at correlations and independence between different noise signal sources. The use of randomised phase Fourier synthesis models was proposed as a potential way of gaining understanding of features in delay space. Some keywords for searching in this area would be: delay space, Takens theorem, L. A. Smith, D. Broomhead, nonlinear systems, embedding dimension, signal reconstruction, *etc.* Dr Lenny Smith

(lenny@maths.ox.ac.uk) heads up a team at Oxford on analysis of time series and has worked on time delay space. He also heads up the Centre for the Analysis of Time Series at LSE.

Validation of any synthesised noise environment signals against samples of real signals would be required to provide credibility. There was discussion of the use of optimised fitting methods to automatically determine parameters of the synthesis from sampled sounds. The benefits of using the MATLAB signal processing and optimisation toolboxes were recognised.

5 Summary and recommendations

Rather than testing a speech recognition system with a lot of background noise that presents it with no real challenge, it is clearly more efficient to be able to generate mainly those noise characteristics that are difficult for the system, and then quickly explore the region of "noise space" in which it begins to fail. Those characteristics may of course vary from system to system, and so some feedback from the testing procedure into the generation of test noise is likely to be necessary in an efficient system.

There would be several benefits if Jomega and Simon Roberts could meet Dr Lenny Smith (mentioned above), and also Dr Irene Moroz and Dr Steve Roberts who work on voice morphing methods. All of them are in Oxford and it is hoped that a suitable meeting can be arranged in the near future.

Addendum

A follow-up meeting was held on 10 May 2002 in Oxford.

- Steve Roberts (sjrob@robots.ox.ac.uk) identified Positive Definite Decomposition methods as one approach that could be used to identify features in the noise, and identify the effective dimensionality of the "noise space". This would then enable one to specify how many "axes" need to be explored.

- Patrick McSharry (mcsharry@maths.ox.ac.uk) also considered how delay space embedding would cope with a process like rain noise, generated by an approximately Poisson process of random impacts of differently sized raindrops. A crucial parameter in this would be how the delay time was chosen relative to the mean time between impacts. As a next step, it was proposed that Jomega provide sample noise data so that the possibility of using these methods might be explored.