Communications, Navigation and Mathematics: Part I – Speech Communications
Edward V. Stansfield CMath FIMA, CSci

This article is based on a public lecture given by the author at the University of Reading on 13 November 1997 to celebrate his appointment as a Visiting Industrial Professor. The material has remained in his files and unpublished until now. The primary aim of the lecture was to demonstrate that the basic mathematics which underpins one technology (e.g. digital speech communications) can often be applied to a quite different technology (e.g. navigation systems). This article follows the basic theme of the lecture, but includes more of the mathematics. This is the first part of the article, Speech Communications, the second part on Navigation will appear in the next issue of Mathematics Today.

1 Telephone evolution

The human voice is an acoustic analogue signal, and until the latter half of the 20th century the plain old telephone system (POTS) was analogue throughout since its invention by Alexander Graham Bell in 1876. This is illustrated in Figure 1. A microphone in the caller’s handset converts the acoustic speech waveform into a continuous electrical signal. This is then transmitted via the analogue switched telephone network to the receiver, where the earpiece in the called party’s handset converts the electrical signal back to an acoustic waveform.

![Figure 1: POTS – plain old telephone system](image)

This analogue telephone network served the world well, but it had one major drawback – unavoidable noise. This is added to the signal at every stage, in the cables, in the amplifiers and in the airwaves. The noise problem was particularly acute with long-distance communications, especially intercontinental. Around the middle of the 20th century a method of coding speech into a digital signal was invented by Alec Harley Reeves. Transmitting a digital signal over long distances has the advantage that it can be regenerated as necessary without any loss of fidelity, thereby avoiding the additive noise problem. This has led to what I will refer to here as pretty amazing new stuff (PANS). This is illustrated in Figure 2. Digital communications systems now exist everywhere throughout the world, from cordless phones in the home, mobile phones on the move and the backbone switched telephone network.

I will address herein just one aspect of digital communications technology, namely the conversion between the electrical analogue of the speech waveform and its digitised version. Most of the energy in a speech waveform lies in the frequency range from 300 to 3,400 Hz, and in the traditional POTS the signal is limited to just this range. When the POTS network was first digitised using pulse-code modulation (PCM), this band limiting was maintained. The signal is sampled at 8,000 Hz and logarithmically quantised to 8 bits, requiring 64 kbit/s per speech channel. With the evolution of mobile phones, the available radio spectrum and power levels would not support very many channels if this bit rate had been used, particularly so in view of its susceptibility to noise in the radio spectrum. Speech compression techniques are used to reduce the data rate and thereby increase the number of channels.

2 Speech production and phonetics

To understand why speech can be compressed to lower data rates, it is enlightening to consider its properties and the physical mechanisms which humans use to produce it. Speech is highly redundant; this comes about as a result of the way that thoughts in the brain are transformed into an acoustic speech waveform. The transformation takes place via nerve signals which activate muscles to move the articulators (see Figure 3). At each stage of this process – from electrical impulses through mechanical movement to acoustic waveform – additional information is added, which is ultimately contained in the speech waveform but is not part of the original thought in the brain.

![Figure 3: Speech production mechanism](image)

The redundancy in speech helps to explain the following phenomenon. Consider a sentence in which all the consonants have been replaced by the plosive consonant /k/, retaining only the vowels: kak kou ukkekkkakk kkik. Now consider the same sentence, but with the vowels replaced by a schwa (neutral) vowel /ə/, retaining the consonants: can yoo understand this. If it is not already clear, then by attempting to pronounce these two sentences, it should become clear that the intelligibility in speech is primarily contained in the consonants – the plosives, the fricatives, the stop consonants, glottal stops and so on. The sentence is of course: can you understand this. This example shows that vowel sounds add little to the basic information content of speech.
The mechanism of speech compression and its digitisation at low data rates is based on a mathematical source-tract model of speech production. Referring to Figure 4, speech is formed by passing air from the lungs through the trachea via the vocal folds into the pharynx, over the tongue and out through the lips. For vowel sounds the vocal folds vibrate at the pitch frequency, which is varied by adjusting the tension in the folds (Figure 5). Fricative sounds are formed by forcing air turbulence at a point of constriction in the vocal tract.

In more detail, speech can be thought of as a sequence of distinct phonemes. Some examples of English phonemes are indicated in Table 1 for vowels and Table 2 for fricatives. Individual phonemes can be described in terms of the position of the tongue, teeth and lips. Other phonemes include stop consonants (plosives) formed by forcing a sudden release of air pressure at a point along the vocal tract. Nasal sounds depend on the coupling between the nasal and vocal tracts by the velum. For a more complete list of phonemes and more details about human speech production, see e.g. Flanagan [1].

If the speech waveform is described in terms of parameters that represent the phonemes, then the data rate required to transmit this information is considerably less than the 64 kbit/s required for PCM. Indeed, if we ignore the speech characteristics that distinguish one person from another, then the information rate is typically less than 100 bit/s. In vocoders used by the military, and in mobile phones widely used in the commercial world, it is these properties of speech that are exploited to reduce the data rate from the 64 kbit/s of PCM to something much less. The degree of compression allowed depends on the reduction in voice quality and individual characteristics that are acceptable for the purpose. Military vocoders can operate at rates lower than 1,200 bit/s, whereas the lowest rate currently used in civilian mobile telephone networks is 6,500 bit/s.

<table>
<thead>
<tr>
<th>Degree of constriction</th>
<th>Tongue hump position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front</td>
<td>Central</td>
</tr>
<tr>
<td>High</td>
<td>/i/</td>
</tr>
<tr>
<td>/e/</td>
<td>/æ/</td>
</tr>
<tr>
<td>/æ/</td>
<td>/æ/</td>
</tr>
<tr>
<td>Medium</td>
<td>/æ/</td>
</tr>
<tr>
<td>Low</td>
<td>/æ/</td>
</tr>
<tr>
<td>Labio-dental</td>
<td>/i/</td>
</tr>
<tr>
<td>Dental</td>
<td>/æ/</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/æ/</td>
</tr>
<tr>
<td>Palatal</td>
<td>/æ/</td>
</tr>
<tr>
<td>Glottal</td>
<td>/æ/</td>
</tr>
<tr>
<td>Table 1: Vowel sounds</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location of constriction</th>
<th>Voiced</th>
<th>Unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labio-dental</td>
<td>/i/</td>
<td>/i/</td>
</tr>
<tr>
<td>Dental</td>
<td>/æ/</td>
<td>/æ/</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/æ/</td>
<td>/æ/</td>
</tr>
<tr>
<td>Palatal</td>
<td>/æ/</td>
<td>/æ/</td>
</tr>
<tr>
<td>Glottal</td>
<td>/æ/</td>
<td>/æ/</td>
</tr>
<tr>
<td>Table 2: Fricative consonant sounds</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 Mathematical model of vocal tract

The basic form of the source-tract model of speech production is shown in Figure 6. The vocal tract is the component which dictates the primary information contained in the speech waveform, and about 95% of the information is contained in the frequency range from 300 to 3,600 Hz. This is the reason why in telephony speech is usually digitised by sampling at 8 kHz, slightly above the Nyquist frequency. This corresponds to an interval of 125 μs between samples. The vocal tract itself is mathematically modelled as a piecewise continuous tube of varying cross-sectional area (Figure 7).

![Figure 4: Sagittal cross section of vocal organs](image1)

![Figure 5: Cross section of larynx](image2)

![Figure 6: Source-tract model of speech production](image3)

![Figure 7: Acoustic tube model of vocal tract](image4)
The impulse response of the vocal tract from the opening of the larynx (glottis) to the lips defines its transfer function. If a single impulse is injected at the glottis, the interval between successive impulses at the lips will be twice the time it takes to travel along one section. Since acoustic waves in the vocal tract travel at approximately 350 m/s, they cover a distance of 4.375 cm in each sampling interval. An adult vocal tract is typically about 17.5 cm long and hence $2 \times 17.5/4.375 = 8$ sections are needed to model a speech waveform sampled at 8 kHz. This means that in principle the vocal tract can be modelled using just 8 coefficients related to the cross-sectional area of each section. To allow for deficiencies in the model, 10 coefficients are normally used.

The acoustic tube model can be represented as a lattice filter as shown in Figure 8. In this diagram, the letter D denotes a delay of 125 µs, corresponding to the 8 kHz sampling frequency. In terms of the cross-sectional areas $A_m$, $m = 0 : 10$, the reflection coefficients are defined by

$$k_m = \frac{A_m - A_{m-1}}{A_m + A_{m-1}}, \quad m = 1 : 10.$$  

One advantage of using reflection coefficients instead of cross-sectional areas is that from a coding point of view stability is guaranteed because they will always be in the range $(-1, +1)$.

![Figure 8: Lattice filter model of vocal tract](image)

The recursive nature of the lattice filter implies that its impulse response has an infinite duration. An equivalent transversal filter model of the lattice filter, which is more suitable for computation, is shown in Figure 9. This model can be interpreted as a linear predictor, since the feedback created by the coefficients $a_m$, $m = 1 : 10$, at each sampling instant, when added to the next excitation sample, generates the next speech sample.

![Figure 9: Transversal filter model of vocal tract](image)

There is a unique mapping from the reflection to the predictor coefficients, given by the recursion equations

$$a_i^{(m)} = a_i^{(m-1)} - k_m a_{i-1}^{(m-1)}$$  

with

$$a_0^{(m)} = -1 \quad \text{and} \quad a_m^{(m)} = k_m$$

for

$$i = 0 : M \quad \text{and} \quad m = 1 : M.$$  

Starting with $m = 1$, the recursion is run for $i = 0 : M$ where $M = 10$ is the predictor order. The process is then repeated for $m = 2$, and so on until $m = M$, at which point $a_i = a_i^{(M)}$, $i = 1 : M$, are the prediction coefficients. The inverse mapping from the predictor to the reflection coefficients is also recursive

$$k_m = a_m^{(m)} \quad \text{and} \quad a_i^{(m-1)} = a_i^{(m)} + k_m a_{i-1}^{(m)} \over 1 - k_m^2$$

with

$$a_0^{(m)} = -1 \quad \text{for} \quad i = 0 : M \quad \text{and} \quad m = 1 : M.$$  

The transformation in this case is initialised with $m = M$ and run for $i = 0 : M$, then repeated for $m = M - 1$, and so on until $m = 1$, at which point all $M$ reflection coefficients $k_m$, $m = 1 : M$, will be available.

### 4 Speech analysis and synthesis

A process used to estimate the coefficients from samples of the speech waveform is based on least squares. Consider again the transversal filter model of the vocal tract shown in Figure 9. Let $e_n$, $n = 1, 2, \ldots$, be the input samples at the glottis and $s_n$, $n = 1, 2, \ldots$, the speech output samples at the lips. The sampled time domain equations for this model are then given by

$$s_n = e_n + \sum_{m=1}^{M} a_m s_{n-m}.$$  

Expanding this into a set of $N \gg M$ simultaneous linear equations, expressed in matrix form, gives

$$\begin{pmatrix} s_n \\ s_{n-1} \\ \vdots \\ s_{n-N+1} \end{pmatrix} = \begin{pmatrix} e_n \\ e_{n-1} \\ \vdots \\ e_{n-N+1} \end{pmatrix} + \begin{pmatrix} s_{n-1} & s_{n-2} & \cdots & s_{n-M} \\ s_{n-2} & s_{n-3} & \cdots & s_{n-M-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n-N} & s_{n-N-1} & \cdots & s_{n-M-N+1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{pmatrix}.$$  

$$\Leftrightarrow S = E + C.A.$$  

The required coefficient vector $A$ is that vector for which the excitation vector $E$ has minimum energy, that is, for which the energy $\varepsilon = E^T E \rightarrow \text{minimum}$. Since the equations are linear in $A$, this is just the least-squares solution. In detail

$$E = S - CA \Rightarrow \varepsilon = E^T E = (S - CA)^T (S - CA)$$

$$\Rightarrow \frac{d \varepsilon}{d A} = -C^T (S - CA).$$

Then for minimum energy we have

$$\frac{d \varepsilon}{d A} = 0 \Rightarrow C^T S = C^T CA \Rightarrow A = (C^T C)^{-1} C^T S.$$  

This shows that the $M \times 1$ coefficient vector $A$ is obtained by first computing the $M \times M$ autocorrelation matrix $C^T C$ of $N + M$ speech samples, taking its inverse, and then multiplying the result by the $M \times 1$ autocorrelation vector $C^T S$ of speech samples.

Figure 10 is a block diagram of a generic digital speech coder/decoder based on linear prediction. In the transmitter the vocal tract prediction coefficients (estimated from the speech input samples) are converted to reflection coefficients, checked for
stability and encoded along with information about the excitation waveform. In the receiver, the coefficients are used directly in a lattice filter driven by a suitable synthetic excitation waveform to produce synthetic speech.

Since the vocal tract shape and excitation change continuously with time, it is necessary to repeat the computations and update the transmitted parameters at regular intervals. This is typically done every 20 ms, a period which is significantly shorter than the duration of individual phonemes. The perceived quality of the synthesised speech depends on the way in which the excitation function is coded and transmitted. Simple pulses for voiced sounds and random noise for unvoiced sounds can provide highly intelligible but unnatural sounding speech at low data rates. More complex schemes to encode the excitation provide more natural speech quality but require higher data rates. The details of how this is achieved are beyond the scope of this article.

**REFERENCES**


---

### On the Number of Sudoku Grids

**Siân K. Jones**† CMath MIMA, **Stephanie Perkins**† CMath MIMA and **Paul A. Roach**† CMath MIMA

The subject of the number of sudoku grids has sparked much recent attention in *Mathematics Today* (R. Franklin (October 2012), Charles Evans (April 2013) and Ian Stewart (June 2013)). This feature addresses known results for the numbers of sudoku grids of different sizes, and explains how a well-formed symmetry group can be used to simplify the enumeration of sudoku grids.

Sudoku is usually thought of as a 9 × 9 puzzle which is further subdivided into mini-grids of size 3 × 3, with each of the 81 cells of the grid to be filled with the digits 1 to 9 such that each digit appears exactly once in each row, column and mini-grid. In fact, although published sudoku puzzles are generally 9 × 9 in size, other dimensions can be used, and for every non-prime dimension n there is an n × n sudoku grid. However, for some sizes of sudoku more than one size of mini-grid can be chosen. For example, a 6 × 6 sudoku (known as sudokuo) can have mini-grids of size 3 × 2 or 2 × 3 (although these are essentially a rotation from one to the other), and a 12 × 12 sudoku can have mini-grids of size either 3 × 4 or 6 × 2 (leading to very different puzzles). Published puzzles show incomplete grids, with a number of cells pre-filled with fixed, or given digits, chosen to ensure that a solution is unique.

Formally a sudoku grid, $S_{x,y}$, is a n × n array subdivided into n mini-grids of size x × y (where n = xy); the values 1,...,n are contained within the array in such a way that each value occurs exactly once in every row, column and mini-grid. We denote a sudoku grid of size n × n with mini-grids of size x × y, where n = xy, as $S_{x,y}(n)$. $S_{x,y}(n)$ and the number of ways of arranging the values in $S_{x,y}(n)$ as $S_{x,y}(n)$, $S_{x,y}(a,b)$ is a specific mini-grid in band a and stack b of $S_{x,y}$ and $[S_{x,y}(a,b)]_{i,j}$ a specific cell in tier i and pillar j of the mini-grid $S_{x,y}(a,b)$ (see Figure 1).

---

† University of South Wales