SEPARATION OF SPEECH SOURCE AND FILTER BY TIME-DOMAIN DECONVOLUTION

C. William Thorpe

National Voice Centre, University of Sydney, Australia

ABSTRACT

A subtractive deconvolution algorithm is described which allows one to separate a voiced speech signal into two components, representing the time-invariant and dynamic parts of the signal respectively. The resulting dynamic component can be encoded at a lower data rate than can the original speech signal. Results are presented which validate the utility of decomposing the speech waveform into these two components, and demonstrate the ability of the algorithm to represent speech signals at a reduced data rate.

1. INTRODUCTION

Speech can be thought of as the result of a convolution between two time-varying components, the glottal source and vocal tract filter, neither of which can be measured directly. Trying to separate out these components is useful because they each contain particular information about aspects of the speech production mechanism. Algorithms for separating these two components rely on some assumption on how the characteristics of the signals differ. For instance, linear prediction presupposes that the vocal tract filter is all-pole whereas the glottal excitation is all-zero [1].

In this paper, voiced speech is modelled as a convolution between a relatively time-invariant component (which may be equated to the average glottal pulse and vocal tract impulse response shape) and a time-varying component which encapsulates all the dynamic features of the speech signal. Because of the disparate nature of the variability of these components, they are able to be separated by means of time-domain deconvolution [2]. Section 2 describes the deconvolution algorithms, while results of their application to the low data-rate coding of speech signals are presented in Section 3.

2. METHODS

It is convenient to separate an utterance $s(t)$ into M contiguous segments $s_m(t)$, so that

$$s(t) = \sum_{m=1}^{M} s_m(t-T_{em}), \qquad \textbf{(1)}$$

where T_{em} is the point of application of the m^{th} segment. Each segment can be further described by the convolution

$$s_m(t) = v_m(t) \odot e_m(t) + c_m(t), \qquad \textbf{(2)}$$

where $\odot$ is the convolution operator, $e_m(t)$ is the excitation signal, and $v_m(t)$ is the vocal tract filter response during the m^{th} segment. The quantity $c_m(t)$ is called the contamination, and embodies all parts of the speech signal which are not described by the convolution term. The excitation $e_m(t)$ is noise-like for unvoiced segments of a speech utterance, and consists of a quasi-periodic train of "glottal pulses" for voiced segments. A further simplification to the model can be obtained by limiting $e_m(t)$ to one of two forms – a noise source $e_u(t)$ for unvoiced segments and a fixed-shape pulse $e_g(t)$ for voiced segments. The implication here is that each segment represents a single cycle of the glottal vibration.

2.1. Shift-and-add Blind Deconvolution

Voiced speech can be thought of as the outcome of applying a series of fixed shape glottal pulses $e_g(t)$ to the time-varying filter $v_m(t)$ of the vocal tract. An equivalent process also exists in astronomical imaging where a fixed image of the stars is filtered by time-varying distortions of the atmosphere [3]. By collecting a series of short time exposures, an ensemble of differently distorted images is obtained.

Under certain conditions where the time-varying filter is sufficiently variable [4], we can extract an estimate of the invariant component (glottal pulse shape or astronomical image respectively) by synchronously averaging the resulting ensemble of short-time segments. The process, termed *shift-and-add* (SAA) in the astronomical literature, entails *shifting* each distorted output such that its brightest component is at the origin, and *adding* across the ensemble to obtain the average. For voiced speech therefore, the estimated invariant glottal pulse shape $s_{sa}(t)$ is given by:

$$s_{sa}(t) = \langle s_m(t+T_m) \rangle_{m_v} \qquad \textbf{(3)}$$

where T_m is the instant where $|s_m(t)|$ is greatest and $\langle\cdot\rangle_{mv}$ denotes ensemble averaging over the voiced segments m_v. Replacing $s_m(t)$ in **(3)** by its expansion **(2)**, with $e_m(t) = e_g(t)$, gives

$$s_{sa}(t) = e_g(t) \odot \langle v_m(t+T_m) \rangle_{m_v} + \langle c_m(t+T_m) \rangle_{m_v} \qquad \textbf{(4)}$$

Under the assumption that $v_m(t)$ varies sufficiently from segment to segment throughout the utterance, $\langle v_m(t+T_m) \rangle$ reduces to an impulse function, and if the contamination is independent of $v_m(t)$ and small enough so that $\langle c_m(t+T_m) \rangle = 0$, the resultant average $s_{sa}(t) \approx e_g(t)$. In practice, $v_m(t)$ does not vary in a completely unbiased manner, so $s_{sa}(t)$ contains contributions from the part of $v_m(t)$ that persists throughout the utterance [5]. Fig.1(a) shows the result of applying this algorithm to an utterance spoken by a male speaker.

2.2 Subtractive Deconvolution - the CLEAN Algorithm

The processing described in the previous section allows us to extract a representation of the invariant component of the voiced speech signal. To estimate the variant component $v(t)$, it is necessary to deconvolve $e_g(t)$ from $s(t)$ in some manner. Care is needed, however, because the additive contamination $c_m(t)$ can be

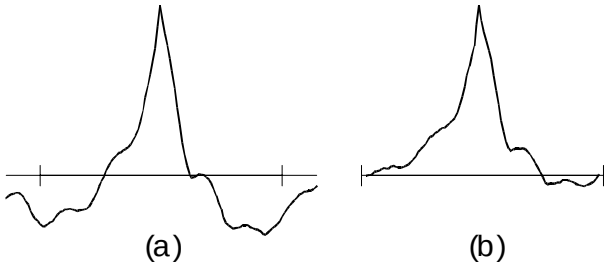

Fig. 1 (a) Result of applying the SAA algorithm to the utterance *When sunlight strikes raindrops in the air, it acts like a prism, and forms a rainbow*, spoken by a male. The marks on the horizontal axis indicate the extent of the average pitch interval 10ms. **(b)** Deconvolution kernel $g(t)$ obtained from the $s_{sa}(t)$ shown in (a) by truncating and adding an offset so that the endpoints approximate zero.

appreciable if the actual excitation for any particular segment differs significantly from $e_g(t)$. The CLEAN algorithm was originally developed in the context of deblurring radio-astronomical images [6], where a "clean" image can be reconstructed from the distorted measurement and the known blurring filter which is best described in the image-domain rather than in the frequency domain. For speech signals likewise, the spectrum of the blurring filter $e_g(t)$ exhibits a large dynamic range which would cause difficulties for deconvolution algorithms that operate in the frequency domain.

Subtractive deconvolution is based on the idea that a filtered signal can be considered as consisting of many copies of the filter's impulse response, each one shifted in time and weighted by the amplitude of the unfiltered signal at that instant [2]. Hence we can recursively estimate the variant component characterising a complete utterance by repeatedly estimating the amplitude and position of each copy of $e_g(t)$ and subtracting it from $s(t)$. Simultaneously, the "clean" unfiltered signal $v(t)$ is constructed by superimposing discrete pulses of appropriate amplitude and position which represent each of the copies of $e_g(t)$.

For a segment of speech the CLEAN subtractive deconvolution algorithm is defined by the following sequence of steps. Initially, $r(t)$ is set equal to $s(t)$, and $v(t)$ equal to 0, for $0 < t < \tau^{seg}$, where τ^{seg} is the duration of the segment. The deconvolution kernel $g(t)$ is set to a modified form of $s_{sa}(t)$, such modification being necessary to ensure that the end-points of $g(t)$ are of zero amplitude. Thereafter, for each iteration labelled by j:

1. The position p_j at which $|r_j(t)|$ is greatest is located.

$$p_j = \underset{t}{\text{argmax}} \; |r_j(t)|, \qquad \textbf{(5)}$$

2. The amplitude υ_j corresponding to the weight of the j^{th} copy of $g(t)$ in $s(t)$ is estimated as

$$\upsilon_j = \gamma r_j(p_j) \big/ g(0) \qquad \textbf{(6)}$$

where γ is the loop gain.

3. The CLEAN signal is updated to reflect this newly estimated copy of $g(t)$:

$$v_j(t) = v_{j-1}(t) + \upsilon_j \delta(t - p_j), \quad 0<t<\tau^{seg}. \qquad \textbf{(7)}$$

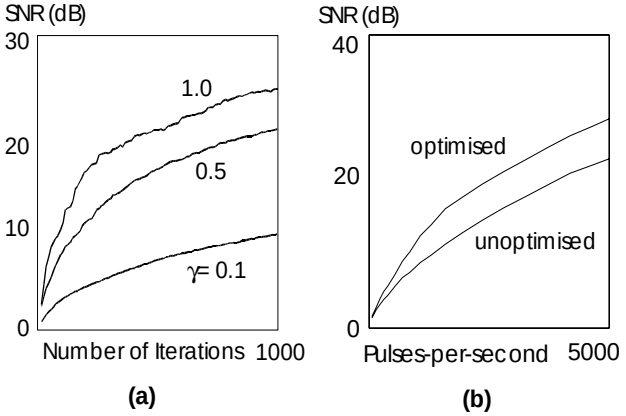

Fig. 2 Signal-to-noise ratio (SNR) obtained with the CLEAN algorithm applied to the segment of speech shown in Fig.3: **(a)** SNR versus the number of iterations of the CLEAN algorithm for three values of the loop gain parameter γ. **(b)** SNR versus the number of non-zero pulses in the CLEAN signal with and without pulse amplitude re-optimisation (γ=0.7).

4. The residual is reduced:

$$r_{j+1}(t) = r_j(t) - \upsilon_j g(t - p_j), \quad p_j + \tau_- < t < p_j + \tau_+ \qquad \textbf{(8)}$$

where $\tau_- < t < \tau_+$ is the interval over which $g(t)$ is non-zero when the maximum magnitude in $g(t)$ occurs at $t=0$.

5. If the pulse position p_j signifies a new distinct pulse in $v(t)$ (i.e. $v_{j-1}(p_j) = 0$), the pulse counter N_p is incremented.

6. Steps 1-5 are repeated until either $\upsilon_j < \eta_{cl}$, $N_p = P_{max}$, or $j \geq J_{max}$, where η_{cl} is a threshold on the largest magnitude in the residual signal $r_j(t)$, P_{max} limits the number of non-zero pulses in the CLEAN signal, and J_{max} limits the maximum number of iterations (a necessary condition in cases where the algorithm does not converge).

As γ is increased from 0, the rate of convergence of the algorithm also increases, until it becomes unstable at a critical value of γ, which depends upon the forms of both $g(t)$ and $s(t)$. Fig.2(a) shows the variation of SNR (the level of the residual signal $r_j(t)$) versus j the number of iterations, for three values of γ, when $g(t)$ and $s(t)$ are as shown in Figs.1(b) and 3 respectively. Fig. 4(a) shows the final CLEAN signal corresponding to Fig. 1(b) and 3, when the SNR is equal to 15dB and γ = 0.5. The residual $r(t)$ is shown in Fig. 4(b). Fig. 4(c) shows the reconstruction of $s(t)$ generated by convolving the CLEAN signal in Fig. 4(a) with the $e_g(t)$ shown in Fig. 1(b).

At each iteration of the CLEAN algorithm, the amplitude of the new pulse is estimated by considering only the peak magnitude of the speech signal and the effects on it of the previously estimated pulses. Later pulses may modify the residual signal such as to make the current estimate of the pulse position or amplitude inaccurate. If a peak in the residual signal reappears in later iterations at the position of a particular pulse, its amplitude is updated. However, significant improvements to the fidelity of the reconstructions can be obtained by *re-optimising* the pulse amplitudes after they have been located by the CLEAN algorithm, without increasing the number of non-zero pulses.

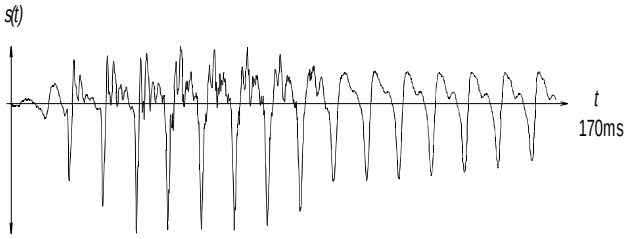

Fig. 3 Segment of speech used to illustrate the CLEAN algorithm.

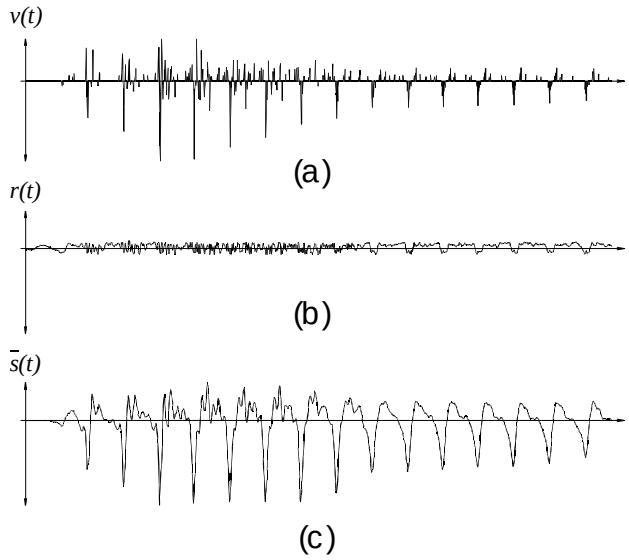

Fig. 4 Results of applying 430 iterations of the CLEAN algorithm to the segment of speech shown in Fig. 3 with a loop gain γ=0.5. **(a)** CLEAN signal consisting of the equivalent of 1580 non-zero pulses per second. **(b)** Residual; and **(c)** reconstructed speech signal. The SNR is 15dB.

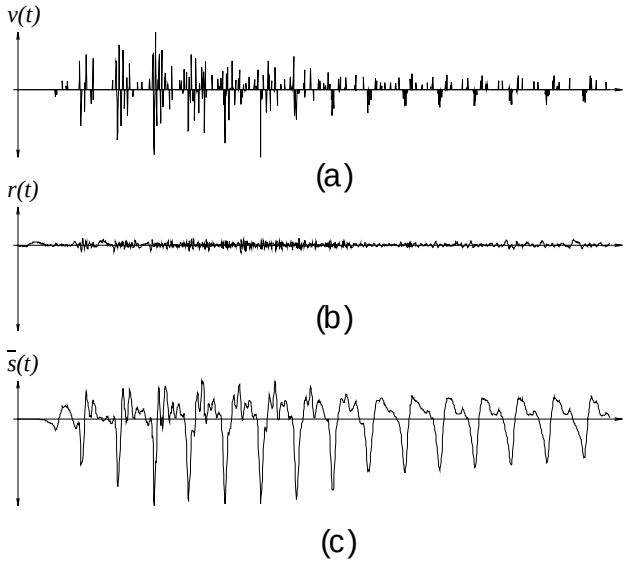

Fig. 5 Re-optimising of the CLEAN pulse amplitudes shown in Fig.4(a). The three traces correspond to those shown in Fig.4, after pulse amplitude optimisation. The SNR is increased to 21dB.

It is convenient to present the optimisation algorithm in the sampled time domain. Readers should bear in mind, however, the implied correspondence between the sample index n and the time index t ($t = nT_s$, where $1/T_s$ is the sampling frequency).

The mean square error E between the original speech $s[n]$ and the speech signal reconstructed from the CLEAN signal is given by

$$E = \sum_{n=-\infty}^{\infty} \left[s[n] - \sum_{j=1}^{N_p} \upsilon_j g[n-p_j] \right]^2 \qquad \textbf{(9)}$$

For computational reasons, it is convenient to perform the amplitude optimisation on blocks of N_{opt} CLEAN pulses, holding the remaining pulses constant. Denoting the index of the first pulse in such a block as j_o, the error E_{opt} within the block becomes

$$E_{\text{opt}} = \sum_{n=-\infty}^{\infty} \left[y[n] - \sum_{j=j_o}^{j_o+N_{\text{opt}}-1} \upsilon_j g[n-p_j] \right]^2 \qquad \textbf{(10)}$$

where

$$\begin{aligned} y[n] &= s[n] - \sum_{i=1}^{j_o-1} \upsilon_i g[n-n_i] - \sum_{i=j_o+N_{\text{opt}}}^{N_p} \upsilon_i g[n-p_i] \\ &= r[n] + \sum_{i=j_o}^{j_o+N_{\text{opt}}-1} \upsilon_i g[n-p_i] \end{aligned} \qquad \textbf{(11)}$$

is the signal remaining when the effects of all the CLEAN pulses that are not within the optimisation block have been removed from the speech signal. Note that it is more convenient to compute $y[n]$ via the second form in **(11)** because $r[n]$ is available from the CLEAN algorithm.

Setting the partial derivatives of E_{opt} with respect to each of the υ_j to zero leads to the matrix equation

$$\sum_{k=j_o}^{N_{\text{opt}}-1} g_{p_k p_j} \upsilon_k = c_{p_j}, \qquad j = j_o \ldots N_{\text{opt}}-1 \qquad \textbf{(12)}$$

where

$$g_{p_i p_j} = \sum_{n=-\infty}^{\infty} g[n-p_i] g[n-p_j] \qquad \textbf{(13)}$$

is the $(p_i+p_j)^{\text{th}}$ term from the autocorrelation of the CLEAN kernel $g[n]$, and

$$c_{p_i} = \sum_{n=-\infty}^{\infty} g[n-p_i] y[n] \qquad \textbf{(14)}$$

is the p_i^{th} term from the cross-correlation between the CLEAN kernel $g[n]$ and the modified speech signal $y[n]$ for the current optimisation block. Standard matrix solving techniques can be invoked to solve **(12)** and thereby obtain "optimised" values for the CLEAN pulses υ_k.

Fig.5(a) shows the CLEAN signal obtained after optimising the CLEAN pulses shown in Fig.4(a). The resulting reconstructed

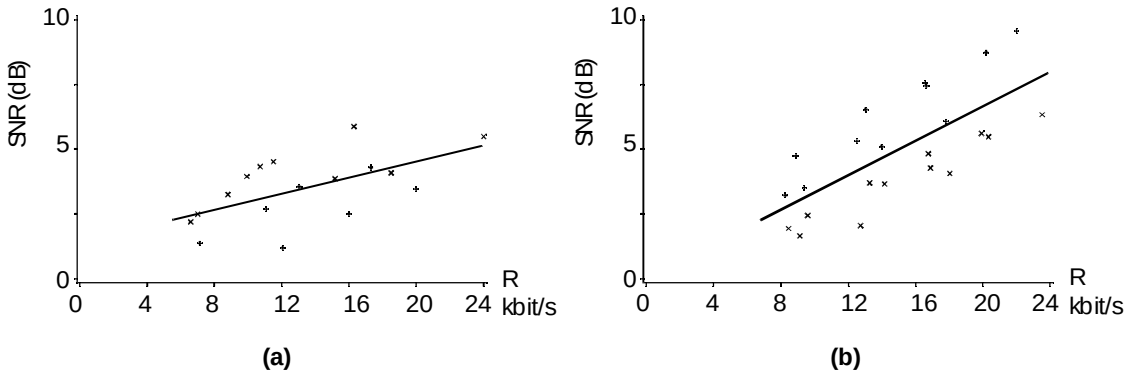

Fig. 6 SNR as a function of data-rate for utterances processed by **(a)** SAA/CLEAN, and **(b)** MP-LPC. Utterances by both male and female speakers are included. The line drawn on each graph is a linear regression through all the points.

speech signal is shown in Fig.5(c). The improvement in SNR over the un-optimised version is indicated by the curves shown in Fig.2(b). These graphs indicate that pulse amplitude re-optimisation provides roughly 6dB improvement to the SNR of the reconstructed speech over the basic CLEAN algorithm.

3. LOW DATA-RATE SPEECH CODING

As indicated by the signals shown in Figs.4(a) and 5(a), the CLEAN signal obtained from a speech utterance consists of non-zero "CLEAN pulses" interspersed with zero-valued samples. Depending on the utterance and the level to which the residual signal is reduced by CLEAN, the number of CLEAN pulses can be as low as 2000 per second whilst still providing good reconstruction (refer to Fig.2(b)). This is considerably less than the number of samples in the speech signal, implying that the CLEAN pulses can be used as a low data rate representation of the speech signal. However, because the CLEAN pulses are not uniformly spaced, both their amplitudes and positions need to be encoded. In order to take full advantage of the benefits implied by the reduced number of pulses, efficient means of encoding the pulse positions and amplitudes must be employed. For illustrative purposes here, the pulse amplitudes are quantised and encoded with 2-4 bits per sample, with run length coding (block lengths of 3-6bits) used to encode the pulse intervals. The SAA signal adds insignificantly to the overall data-rate.

Fig.6 shows the SNR of speech that has been processed at various data rates by either SAA/CLEAN or multi-pulse LPC (MP-LPC). Results are shown for speech uttered by both male and female speakers. These graphs indicate that at a data rate of 12kbit/s the two schemes provide reconstructed speech with similar SNR. However, the SNR of the MP-LPC reconstructions increases much more rapidly than that of the SAA/CLEAN reconstructions with higher data rates.

4. CONCLUSIONS

This paper has presented a new, straightforward and simple method for analysing a speech record, encoding it for economical storage and transmission, and resynthesising it. The CLEAN algorithm has some similarities with the MP-LPC pulse identification algorithm, particularly with regard to the optimisation of the pulse amplitudes. However, here the computationally simple SAA technique is employed to extract a long-term "filter" from the speech signal, rather than a time-varying LPC filter as in MP-LPC. The results reported herein confirm that this scheme provides a similar performance to the MP-LPC technique at data rates of about 12kbit/s.

5. REFERENCES

1. Markel JD and Gray, Jr. AH, *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976
2. Bates JHT, McKinnon AE, and Bates RHT, "Subtractive image reconstruction. I: Basic theory", *Optik*, **61**: 349-364, 1982.
3. Cady FM and Bates RHT, "Speckle processing gives diffraction-limited true images from severely aberrated instruments", *Optics Letters*, 5:438-440, 1980
4. Bates RHT, "Astronomical speckle imaging", *Physics Reports*, **90**:203-297, 1982,
5. Brieseman NP, Thorpe CW, and Bates RHT, "Nontactile estimation of glottal excitation characteristics of voiced speech", *IEE Proc. Pt.A*, **134**:807-813, 1987,
6. Högbom JA, "Aperture synthesis with a non-regular distribution of interferometer baselines", *Astronomical Astrophysics Supplement*, **15**:417-426, 1974