A PRACTICAL MULTICHANNEL DEREVERBERATION ALGORITHM USING MULTICHANNEL DYPSA AND SPATIOTEMPORAL AVERAGING

Mark R. P. Thomas, Nikolay D. Gaubitch, Jon Gudnason and Patrick A. Naylor

Imperial College London
Exhibition Road, London SW7 2AZ, UK
E-mail: {mrt102, ndg, jg, p.naylor}@imperial.ac.uk

ABSTRACT

Speech signals for hands-free telecommunication applications are received by one or more microphones placed at some distance from the talker. In an office environment, for example, unwanted signals such as reverberation and background noise from computers and other talkers will degrade the quality of the received signal. These unwanted components have an adverse effect upon speech processing algorithms and impair intelligibility. This paper demonstrates the use of the Multichannel DYPSA algorithm to identify glottal closure instants (GCIs) from noisy, reverberant speech. Using the estimated GCIs, a spatiotemporal averaging technique is applied to attenuate the unwanted components. Experiments with a microphone array demonstrate the dereverberation and noise suppression of the spatiotemporal averaging method, showing up to a 5 dB improvement in segmental SNR and 0.33 in normalized Bark spectral distortion score.

1. INTRODUCTION

Dereverberation and noise suppression play an important role in speech signal processing. Reverberation components impair the intelligibility of a speech signal and have an adverse effect upon processing algorithms such as recognition and classification. Noise from computer fans, air ducting and other talkers can have equally undesirable consequences. A common means of attenuating these unwanted signals is beamforming, applied to an array of microphones, using the spatial diversity of room transfer functions and noise sources to attenuate the unwanted reverberation and noise components.

Beamforming is a type of *spatial averaging* which produces the greatest enhancement when the wanted components display significantly more interchannel correlation than the unwanted components. This is generally not the case for distant reflections (whose interchannel delay is low) and acoustic noise sources, so a more sophisticated algorithm is required for further enhancement. The quasi-periodicity of voiced speech can be used as a basis for *spatiotemporal averaging* [1]. By averaging the LP residuals [2] over neighbouring larynx cycles from a delay-and-sum beamformer (DSB), the true residual is reinforced and temporally uncorrelated reverberation and noise components are attenuated. LP synthesis with the processed residual gives a cleaner speech signal. The algorithm also uses periods of voiced speech to determine an equalisation filter [3] which performs the equivalent operation of temporal averaging for both voiced and unvoiced speech, further reducing reverberation and noise.

For spatiotemporal averaging to function, an accurate estimation of glottal closure instances (GCIs) is required. The Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) [4] accurately estimates GCIs from clean speech and the multichannel extension [5] exploits the spatial diversity of room transfer functions to give accurate GCI estimation in reverberant environments. The final component for dereverberation by spatiotemporal averaging is voiced/unvoiced/silence detection. Multichannel DYPSA (MC-DYPSA) searches for GCIs in any signal, resulting in spurious GCI candidates during unvoiced speech, so the spatiotemporal averaging algorithm must know during which periods to apply temporal averaging or an equalisation filter alone. A detection algorithm [6] combines a series of metrics to estimate the probability of a frame being voiced, unvoiced or silence.

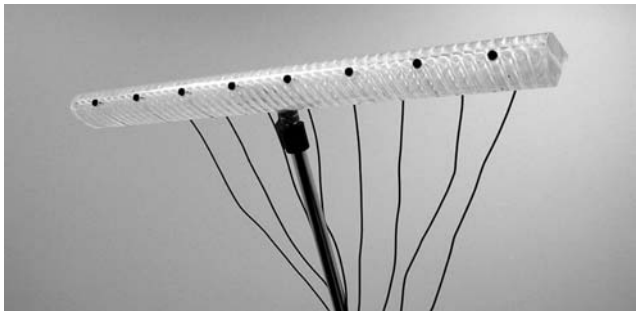

Figure 1: Microphone array comprising eight AKG C417 microphones are placed at 50 mm intervals.

Dereverberation methods can be split into three main categories: (i) beamforming (ii) speech enhancement and (iii) blind channel estimation/equalizaiton. Several existing algorithms are reviewed in [7]. The key contribution of the current paper is to combine the methods described above into a practical (online) and computationally efficient speech enhancement algorithm, which does not require knowledge of the room transfer functions and to demonstrate its applicability in real environments. The proposed method is evaluated with multichannel recordings, captured with a custom microphone array (Figure 1). The remainder of the paper is organised as follows. Section 2 formulates the problem. Section 3 discusses the algorithm in detail. Test results are presented in section 4 and conclusions are drawn in section 5.

2. PROBLEM FORMULATION

Consider a speech signal $s(n)$ produced in a reverberant environment, received by an array of M microphones, through a channel $h_m(n)$ from the source to microphone m. The received signal at

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

Authorized licensed use limited to: Imperial College London. Downloaded on July 15, 2009 at 12:56 from IEEE Xplore. Restrictions apply.

microphone m is $x_m(n) = h_m(n) * s(n)$, where $*$ denotes linear convolution. The aim is to estimate an enhanced speech signal, $\hat{s}(n)$ from the $x_m(n)$. $m = 1.2. .... M$.

LP analysis [2] describes a speech signal as a linear combination of p past samples, such that

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e(n) \tag{1}$$

where a_k are the clean LP coefficients and $e(n)$ is the clean LP residual. Similarly, LP analysis can be applied to each microphone output

$$x_m(n) = -\sum_{k=1}^{p} b_{m,k} x_m(n-k) + e_m(n) \tag{2}$$

where $b_{m,k}$ are the LP coefficients for channel m and $e_m(n)$ is the corresponding LP residual. A single set of best-fit LP coefficients, b_k, may be found with multichannel LPC analysis which closely match a_k. This analysis is presented in detail in [8].

In order to obtain the enhanced speech signal, $\hat{s}(n)$, an enhanced LP residual, $\hat{e}(n)$, is obtained from $e_m(n)$ by spatiotemporal averaging. LP synthesis then resynthesizes the speech signal

$$\hat{s}(n) = -\sum_{k=1}^{p} b_k \hat{s}(n-k) + \hat{e}(n) \tag{3}$$

3. THE ALGORITHM

The algorithm comprises four parts: time-delay-of-arrival (TDoA) estimation with GCC-PHAT, voiced/unvoiced/silence detection, GCI detection with Multichannel DYPSA and spatiotemporal averaging.

3.1. TDoA Estimation

Both MC-DYPSA and spatiotemporal averaging rely on the correct inter-channel time alignment to maximise the correlation of the direct-path signal across channels. The Generalized Cross-Correlation Phase Transform (GCC-PHAT) [9] is a simple and sufficiently accurate method for the estimation of delay between two channels from moderately reverberant speech signals.

Let the reference channel be $x_{ref}(t)$ and the measurement channel $x_m(t)$. The delay estimate, $\hat{\tau}_{GCC}$, is determined by maximising the cross-correlation between channels

$$\hat{\tau}_{GCC} = \arg\max_{\tau} R_{x_{ref}x_m}(\tau) \tag{4}$$

$$R_{x_{ref}x_m}(\tau) = \int_{\infty}^{\infty} \frac{X_{ref}(\omega) X_m^*(\omega)}{|X_{ref}(\omega)||X_m^*(\omega)|} e^{j\omega\tau} \mathrm{d}\omega \tag{5}$$

where * denotes a complex conjugate operation and R is a weighted inverse Fourier transform of the signal cross-spectra.

The GCC-PHAT method has been shown to be accurate enough for moderate reverberation although it is suboptimal under ideal conditions as it places equal weighting on each frequency [9]. The process is repeated for $M-1$ pairs of microphones to determine the inter-channel delay between microphone 0 and microphone m.

3.2. Voiced/unvoiced/silence Detection

Voiced/unvoiced/silence detection is performed on a speech signal which has been processed with a delay-and-sum beamformer (DSB). The output of the DSB, $\bar{x}(n)$, is found by

$$\bar{x}(n) = \frac{1}{M} \sum_{m=1}^{M} x_m(n - \tau_m) \tag{6}$$

where τ_m is a delay to compensate for the propagation time of channel m and is by TDoA estimation as discussed in Section 3.1.

Voiced segments are determined using a voiced-unvoiced-silence detector based on five measurements [6]: 1) zero crossing rate, 2) energy, 3) autocorrelation coefficient, 4) the first LP coefficient and 5) normalized prediction error (in dB). Each measure is computed over 32 ms frames with 60% overlap, forming a sequence of feature vectors. These vectors are then clustered using an unsupervised EM algorithm [10]. The three clusters are labelled as silence, unvoiced and voiced according to their mean vectors and variances. The unvoiced cluster is chosen to be the one with an autocorrelation coefficient closest to zero mean and 0.5 variance. Of the remaining two clusters, the one with greatest energy is chosen to be voiced. Every vector in the sequence is then evaluated under each of the three Gaussians and classified according to which cluster produces the highest likelihood.

3.3. Multichannel DYPSA

The DYPSA [4] GCI detection algorithm comprises three main parts:

(i) *Group Delay Function* – defined as the average slope of the unwrapped phase spectrum of the short time Fourier transform of the prediction residual. GCI candidates are selected based on the negative-going zero crossings of the group delay function.

(ii) *Phase-Slope Projection* – introduced to generate GCI candidates when a local maximum is followed by a local minimum without crossing a zero. The midpoint between these is identified and projected onto the time axis with unit slope. In this way, GCIs whose negative-going slope does not cross the zero point (those missed by the group delay function) are identified.

(iii) *Dynamic Programming* – uses known characteristics of voiced speech (such as pitch consistency and waveform similarity) and forms a cost function to select a subset of the GCI candidates which are most likely to correspond to the true ones. The subset of candidates is selected according to the minimisation problem defined as

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \boldsymbol{\lambda}^T \mathbf{c}_{\Omega}(r). \tag{7}$$

where Ω is a subset of GCIs of size $|\Omega|$, $\boldsymbol{\lambda}$ is a vector of weighting factors and $\mathbf{c}_{\Omega}(r)$ is a vector of cost elements evaluated at the rth GCI of the subset.

Multichannel DYPSA was proposed in [5] to exploit the spatial diversity of room transfer functions [11]. When the channels are time-aligned, the direct-path signal is common to all channels but reverberation components are less likely to show correlation. MC-DYPSA applies parts (i) and (ii) above to each channel independently and creates an additional cost element based upon the interchannel correlation, penalizing those which occur in a small number of channels and encouraging those in close temporal proximity across channels. This is passed to the Dynamic Programming stage and the most likey GCIs, at sample instants n_l, are

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

Authorized licensed use limited to: Imperial College London. Downloaded on July 15, 2009 at 12:56 from IEEE Xplore. Restrictions apply.

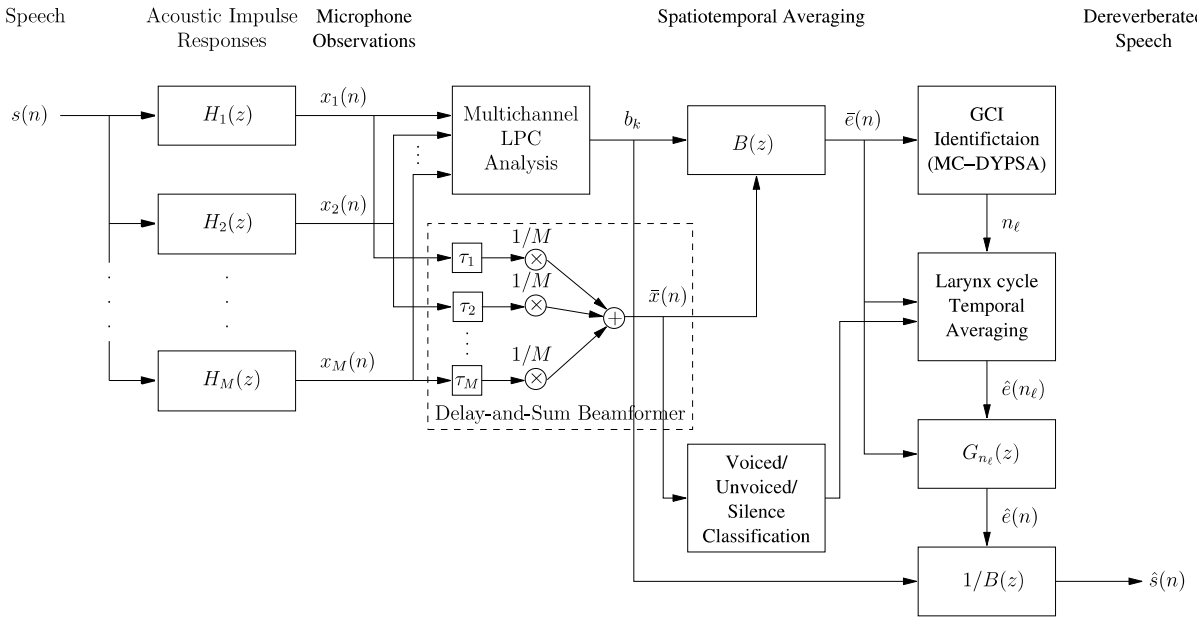

Figure 2: System diagram of the multichannel dereverberation algorithm.

identified. Our experiments have shown that GCI estimation from a reverberant speech signal for $T_{60} = 500$ ms is on average 16% more accurate with MC-DYPSA than single-channel DYPSA applied to an 8-channel DSB and 29% more accurate than DYPSA on a single channel, providing 83% accuracy [5].

3.4. Spatiotemporal Averaging

The DSB prediction residual, $\bar{e}_n$, found by inverse filtering $\bar{x}(n)$ with b_k [2], contains peaks due to GCIs and many spurious peaks due to reverberation and noise. Spurious peaks are uncorrelated among consecutive larynx cycles. Conversely, the main features of prediction residuals from clean speech vary little between neighbouring cycles because of the quasi-stationarity of voiced speech. Performing a weighted average of I neighbouring residuals from larynx cycles of length L of noisy, reverberant speech reinforces the clean speech excitation and attenuates the uncorrelated spurious peaks using

$$\hat{\mathbf{e}}_l = (\mathbf{I} - \mathbf{W})\bar{\mathbf{e}}_l + \frac{1}{2I}\sum_{i=-I}^{I} \mathbf{W}\bar{\mathbf{e}}_{l+i} \tag{8}$$

where $\bar{\mathbf{e}}_l = [\bar{e}(n_l)\ \bar{e}(n_l+1)\ \ldots\ \bar{e}(n_l+L-1)]^T$ is the lth larynx cycle at the output of the DSB with GCI at time n_l, $\hat{\mathbf{e}}_l = [\hat{e}(n_l)\ \hat{e}(n_l+1)\ \ldots\ \hat{e}(n_l+L-1)]^T$ is the lth layrnx cycle of the enhanced residual and $\mathbf{I}$ is the identity matrix. $\mathbf{W} = diag\{\omega_0\ \omega_1\ \ldots\ \omega_{L-1}\}$ is a diagonal weighting matrix to exclude glottal excitation based on the Tukey window [12].

This process can only be applied to segments of voiced speech, leaving reverberation components unaffected on unvoiced speech and silence. Furthermore, in the case of an erroneous GCI, the algorithm will produce incorrect results. To improve robustness, an L-tap equalisation filter $g_l = [g_{l,0}\ g_{l,1}\ \ldots\ g_{l,L-1}]$ for the lth larynx cycle is defined which performs the equivalent operation of temporal averaging. A least squares estimate of g_l is found from $\hat{\mathbf{g}}_l = \min_{\mathbf{g}_l} \|\mathbf{g}_l^T\bar{\mathbf{e}} - \hat{e}(n_l)\|^2$ and is used to update a slowly varying filter

$$\hat{\mathbf{g}}(n_l) = \gamma\hat{\mathbf{g}}(n_{l-1}) + (1-\gamma)\hat{\mathbf{g}}_l \tag{9}$$

where $0 \leq \gamma \leq 1$ is a forgetting factor with typical values in the range $\{0.1 - 0.3\}$, initialised to $\hat{\mathbf{g}}(0) = [1\ 0\ \ldots\ 0]^T$. It is updated only during voiced speech, with the last iteration used for periods of unvoiced speech or silence.

4. RESULTS

The microphone array shown in Fig. 1, consisting of eight AKG C417 microphones spaced linearly at 5 cm intervals, was placed in a 3.3x2.9x2.9 m room with reverberation time (T_{60}) of 0.3 s. A channel estimation was made for each microphone using the MLS method [11]. Utterances of the sentence, "George made the girl measure a good blue vase," by five male and five female talkers were taken from the APLAWD database [13] and played through a loudspeaker at distances 0.5 to 2 m from the microphone array.

The MLS-derived channel estimates were truncated to determine a direct-path impulse response, $h_d(n)$, which was convolved with the clean speech signal to align the unprocessed and processed signals, denoted $s'(n) = h_d(n) * s(n)$. Recording and channel alignment were made at a sampling frequency of $f_s = 48$ kHz. The remainder of the processing was performed at $f_s = 16$ kHz and with the samples high-pass filtered at 100 Hz. The recorded, DSB and spatiotemporal averaged speech samples were evaluated against the clean samples using the segmental Signal-to-Noise Ratio (SNR) [7] and Bark Spectral Distortion (BSD) [14] using 30 ms frames with 50% overlap. The definition of 'noise' in this case is the combination of both reveberation and background noise.

The segmental SNR results, averaged over all ten talkers in APLAWD, are shown in Fig. 3 for (a) reverberant speech at the microphone closest to the talker, (b) DSB speech and (c) spatiotemporal averaged speech. Corresponding BSD results are shown in

Authorized licensed use limited to: Imperial College London. Downloaded on July 15, 2009 at 12:56 from IEEE Xplore. Restrictions apply.

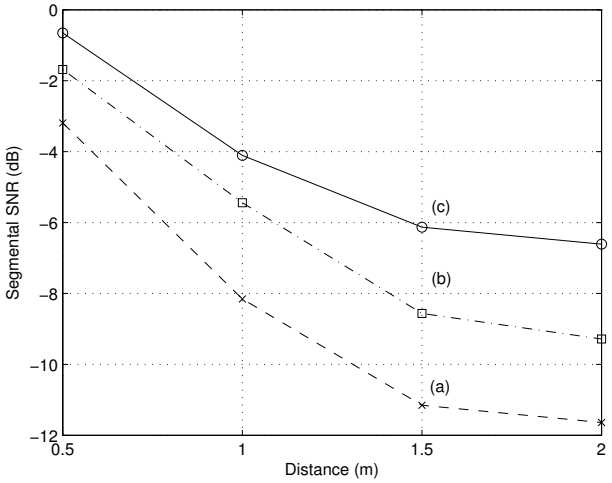

Figure 3: Segmental SNR vs. distance for (a) reverberant, (b) DSB processed and (c) Spatiotemporal averaged speech.

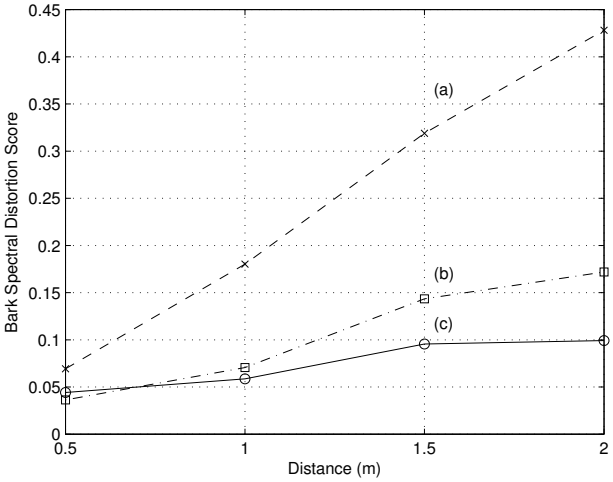

Figure 4: BSD vs. distance for (a) reverberant, (b) DSB processed and (c) Spatiotemporal averaged speech.

Fig. 4. Reverberation and noise reduction of up to 5.0 dB and 0.33 in BSD score are observed at a distance of 2 m, corresponding to 2.7 dB and 0.07 over the DSB. Perceptually, reverberation effects are reduced and the talker appears to be closer to the microphone. The results show a strong correlation with the simulations in [3]. Examples of clean and processed samples can be found at [15].

5. CONCLUSIONS

We have proposed a practical method to exploit the spatial and temporal characteristics of noisy, reverberant speech to attenuate the unwanted signal components. This spatiotemporal averaging algorithm relies on good estimation of GCIs, which are accurately identified with MC-DYPSA. Clean speech samples, played through a speaker and recorded in an office environment, show that significant enhancement in terms of segmental SNR and BSD can be achieved with the proposed algorithm.

6. REFERENCES

[1] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Mulit-microphone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 809–812.

[2] J. Makhoul, "Linear Prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[3] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *Int. Conf. Digital Signal Processing*, Cardiff, UK, July 2007.

[4] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, Language Processing*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[5] M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Multichannel DYPSA for estimation of glottal closure instants in reverberant speech," in *Proc. European Signal Processing Conf. (EUSIPCO) (to appear)*, Poznan, Poland, Sept. 2007.

[6] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 3, pp. 201–212, June 1976.

[7] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Eindhoven, The Netherlands, Sept. 2005.

[8] N. D. Gaubitch, D. B. Ward, and P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4031–4039, Dec. 2006.

[9] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[11] H. Kuttruff, *Room Acoustics*, 4th ed. Taylor & Francis, Oct. 2000.

[12] F. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[13] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Univ. College London, London, U.K., Tech. Rep., June 1987.

[14] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *Selected Areas in Communications, IEEE Journal on*, vol. 10, no. 5, pp. 819–829, Jun 1992.

[15] M. R. P. Thomas, "Samples," May 2007. [Online]. Available: www.commsp.ee.ic.ac.uk/~mrt102/downloads.htm

Authorized licensed use limited to: Imperial College London. Downloaded on July 15, 2009 at 12:56 from IEEE Xplore. Restrictions apply.