

Multimicrophone speech dereverberation using spatiotemporal and spectral processing

Nikolay D. Gaubitch*, Emanuël A. P. Habets[†] and Patrick A. Naylor*

*Dept. Electrical and Electronic Engineering, Imperial College London, UK

Email: {ndg,p.naylor}@imperial.ac.uk

[†]School of Engineering, Bar-Ilan University, Ramat-Gan, Israel

Email: habetse@eng.biu.ac.il

Abstract—Speech signals acquired in a reverberant room with microphones positioned at a distance from the talker are degraded in quality due to reverberation and measurement noise. Therefore, enhancement of reverberant speech is important in hands-free telecommunications applications. The perceptual effects of reverberation can be linked to the room impulse response (RIR) between the talker and the microphone and are characterized by: (i) colouration, due to the strong early reflections and (ii) a distant ‘echoey’ quality due to the decaying tail of the RIR. Accordingly, we present a two-stage multimicrophone method for speech dereverberation. First, spatiotemporal averaging is performed on the linear prediction residual, which primarily reduces the effects of the early reflections. Secondly, a spectral subtraction method is employed to reduce late reverberation. Simulation results with measured RIRs and additive white Gaussian noise illustrate the performance of this method and show that the combined approach performs better than each of the two stages individually.

I. INTRODUCTION

Consider a speech signal, $s(n)$, produced in a reverberant room at a distance from an array of M microphones. The signal observed at the m th microphone can be modelled as the convolution of $s(n)$ with the room impulse response (RIR) between the talker and the m th microphone, $\mathbf{h}_m = [h_{m,0} \ h_{m,1} \ \dots \ h_{m,L-1}]$, and some additive measurement noise $\nu_m(n)$ such that

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}(n) + \nu_m(n), \quad (1)$$

where $\mathbf{s}(n) = [s(n) \ s(n-1) \ \dots \ s(n-L+1)]$ and L is the length of the RIR.

The reverberant and noisy signal $x_m(n)$ is degraded in quality. The severity of degradation depends on the room characteristics, the distance between the talker and the microphones and the level of noise. Reverberation can lead to, for example, reduced intelligibility of the observed speech and lowered performance of speech recognizers. Therefore, speech dereverberation is essential in many hands-free telecommunications applications.

The objective of speech dereverberation is to find an estimate, $\hat{s}(n)$, of the original speech signal, $s(n)$, using the reverberant and noisy observations $x_m(n)$. Several studies of dereverberation have appeared in the literature [1] utilizing, for example, blind system identification [2], multichannel linear prediction [3], homomorphic equalization [4] and prediction residual processing [5], [6]. One common feature of these

methods is that they consider the complete RIR up to the modelling order.

Alternatively, the RIR $\mathbf{h}_m$ can be decomposed into early reflections, $\mathbf{h}_{e,m}$, and late reverberation, $\mathbf{h}_{l,m}$, such that $\mathbf{h}_m = \mathbf{h}_{e,m} + \mathbf{h}_{l,m}$. These two RIR components have distinct perceptual effects on the speech signal [7]: early reflections result in colouration, while late reverberation causes a distant, ‘echoey’ quality. The observed speech signal can also be written as the sum of an early reverberant component, $x_{e,m}(n)$ and a late reverberant component, $x_{l,m}(n)$ such that

$$x_m(n) = x_{e,m}(n) + x_{l,m}(n). \quad (2)$$

Accordingly, in this paper we present a two-stage multimicrophone approach where each stage specifically targets each of these two regions of the RIR. First, we employ the Spatiotemporal Averaging Method for the Enhancement of Reverberant Speech (SMERSH) [8], which primarily reduces the strong early reflections. Secondly, a spectrum estimation and subtraction technique [9] is utilized to attenuate the effect of the reverberation tail. A single microphone method that employs similar philosophy for dereverberation is presented in [10].

The remainder of the paper is organized as follows. In Section II, spectral, temporal and spatial processing is described. Section III discusses the evaluation methodology and presents simulation results. Finally, conclusions from this work are drawn in Section IV.

II. DEREVERBERATION ALGORITHM

In this section, we present the two building blocks of the speech dereverberation algorithm: spatiotemporal averaging [8] and spectral subtraction [9].

A. Spatiotemporal Averaging (SMERSH)

Consider a speech signal expressed in terms of a linear predictor (LP) with a set of LP coefficients and a prediction residual. The main effects of reverberation on the LP of speech reside in the prediction residual, particularly when multiple observations are available [11]. Reverberation affects the LP residual by introducing random peaks of similar strength to the periodic peaks representing the glottal closure instances (GCIs) in clean speech. Furthermore, it has been observed [8] that the LP residual of spatially averaged speech attenuates

these erroneous peaks and allows the strong periodic peaks representing the GCIs to be identified using, for example, the DYPSA algorithm [12]. However, it also contains other erroneous random peaks that are left unattenuated after the spatial averaging. These are uncorrelated among consecutive larynx cycles due to the quasi-periodic nature of voiced excitation. The main features between consecutive larynx cycles in the clean speech LP residual change slowly and show high inter-cycle correlation [13]. Consequently, a moving average operation is applied on neighbouring larynx cycles in voiced speech to suppress the uncorrelated features and, hence, enhance the LP residual. Peaks attributed to GCIs are important to speech quality [14] and are excluded from the averaging process by applying a suitable weight function.

The spatially averaged speech is obtained by applying a delay-and-sum beamformer to the observations $x_m(n)$ [15]

$$\bar{x}(n) = \frac{1}{M} \sum_{m=1}^M x_m(n - \tau_m), \quad (3)$$

where τ_m is a delay to compensate for the propagation time between the source and the m th microphone, and is assumed here to be known.

Subsequently, each enhanced larynx cycle in a voiced speech segment is obtained by averaging the current weighted larynx cycle frame under consideration with $2\mathcal{I}$ of its neighbouring weighted larynx cycles. The result is then added to the original larynx cycle weighted with the inverse weighting function. The ℓ th enhanced larynx cycle is found by

$$\hat{\mathbf{e}}_\ell = (\mathbf{I} - \mathbf{W})\bar{\mathbf{e}}_\ell + \frac{1}{2\mathcal{I} + 1} \sum_{i=-\mathcal{I}}^{\mathcal{I}} \mathbf{W}\bar{\mathbf{e}}_{\ell+i}, \quad (4)$$

where $\bar{\mathbf{e}}_\ell = [\bar{e}(n_\ell) \ \bar{e}(n_\ell + 1) \ \dots \ \bar{e}(n_\ell + \mathcal{L}_\ell - 1)]^T$ is the ℓ th larynx-cycle at the output of the beamformer with its GCI at time n_ℓ , $\hat{\mathbf{e}}_\ell = [\hat{e}(n_\ell) \ \hat{e}(n_\ell + 1) \ \dots \ \hat{e}(n_\ell + \mathcal{L}_\ell - 1)]^T$ is the ℓ th larynx cycle of the enhanced residual, $\mathbf{I}$ is the identity matrix and $\mathbf{W} = \text{diag}\{w_0 \ w_1 \ \dots \ w_{\mathcal{L}_\ell-1}\}$ is a diagonal weighting matrix with the coefficients calculated from the time domain Tukey window [8]. Since the larynx cycles are not strictly periodic but may vary to within a few samples, $\mathcal{L}_\ell$ is set to equal the length of the larynx cycle being processed. Other larynx cycles used in the averaging that have fewer than $\mathcal{L}_\ell$ samples are padded with zeros while those with more than $\mathcal{L}_\ell$ samples are truncated. The parameter $\mathcal{I}$ is important and controls the smoothness of the temporal averaging. Generally, suitable values have been found to lie in the range $1 \leq \mathcal{I} \leq 4$. For the experiments in Section III, we use $\mathcal{I} = 3$.

Spatiotemporal averaging attenuates the reverberant components in the voiced segments of the prediction residual. In order to address unvoiced speech and to take advantage of past accurately identified larynx cycles, an L_i -tap FIR filter is introduced with coefficients $\mathbf{g}_\ell = [g_{\ell,0} \ g_{\ell,1} \ \dots \ g_{\ell,L_i-1}]$. This filter performs the equivalent operation of the inter-cycle averaging. A least squares estimate of $\mathbf{g}_\ell$ is found from $\hat{\mathbf{g}}_\ell = \min_{\mathbf{g}_\ell} \|\mathbf{g}_\ell^T \bar{\mathbf{e}}_\ell - \hat{\mathbf{e}}(n_\ell)\|^2$ and is used to update a slowly

varying filter

$$\hat{\mathbf{g}}(n_\ell) = \gamma \hat{\mathbf{g}}(n_{\ell-1}) + (1 - \gamma) \hat{\mathbf{g}}_\ell, \quad (5)$$

where $0 \leq \gamma \leq 1$ is a forgetting factor with typical values in the range 0.1 – 0.3. The filter is initialized to $\hat{\mathbf{g}}(0) = [1 \ 0 \ \dots \ 0]^T$ with the update performed only during voiced speech segments; in unvoiced speech or silence it is applied at its last update.

It will be demonstrated in Section III that SMERSH primarily attenuates the early reverberation and it inherently suppresses temporally uncorrelated noise. Consequently, we obtain an estimate of the spatially averaged late reverberant signal $\tilde{x}(n) = \hat{s}(n) + \bar{x}_1(n)$, where $\bar{x}_1(n)$ represents the remaining reverberation.

B. Spectral Subtraction

The spectral subtraction method assumes a statistical model of the RIR. According to this model the RIR can be described as a non-stationary process given by

$$h_n = \begin{cases} b(n)e^{-\delta n} & \text{for } n \geq 0 \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where $b(n)$ is a stationary zero mean white Gaussian noise sequence and $\delta = 3 \ln(10)/(T_{60}f_s)$ is the room damping constant, governed by the reverberation time, T_{60} , and the sampling frequency, f_s . This model is valid only when the energy of the direct path is much smaller than the energy of all reflections. This issue has been addressed in [9], where the RIR model is generalized by considering the direct component and the reflections separately. Using (6), it can be shown that there is an additive property between the short-term power spectral density (PSD) of the early speech component and the late reverberant speech component. Hence, the direct component, $\hat{s}(n)$, can be obtained by estimating and subtracting the late reverberant PSD.

The short-time Fourier transform (STFT) is applied to the output signal from SMERSH to obtain $\tilde{X}(i, k) = \hat{S}(i, k) + \tilde{X}_1(i, k)$, which is the i th time frame and k th frequency bin in the time-frequency representation of $\tilde{x}(n)$. The late reverberant PSD can be estimated using [9]

$$\lambda_l(i, k) = e^{-2\delta(k)(n_r - R)} \lambda_r(i - n_r + 1, k), \quad (7)$$

with

$$\lambda_r(i, k) = e^{-2\delta(k)R} \left(\kappa(k) \hat{\lambda}_x(i, k - 1) + (1 - \kappa) \lambda_r(i, k - 1) \right), \quad (8)$$

and

$$\hat{\lambda}_x(i, k) = \beta(k) \hat{\lambda}_x(i - 1, k) + (1 - \beta(k)) |\tilde{X}(i, k)|^2, \quad (9)$$

where $\lambda_r(i, k)$ and $\hat{\lambda}_x(i, k)$ are, respectively, the PSDs of the reverberant speech component, $\tilde{x}_1(n)$, and the input signal, $\tilde{x}(n)$, R denotes the number of samples separating two successive STFT frames. The onset time for the late reverberation, n_r , is typically chosen in the range 30 – 50 ms. The forgetting factor $\beta(k)$ is selected in proportion with the room damping constant [9] and $\kappa(k)$ is a constant, inversely proportional to

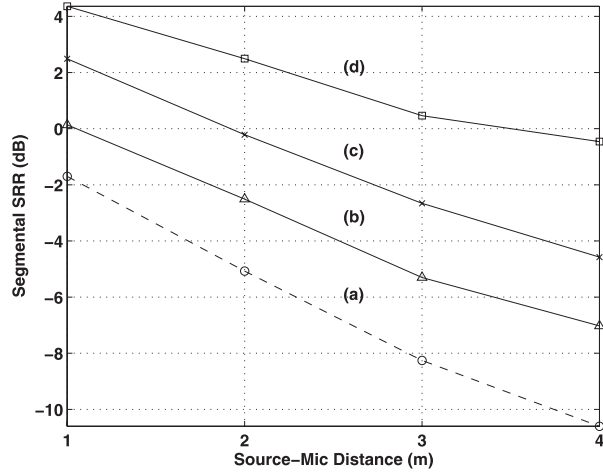


Fig. 1. Segmental SRR for (a) reverberant speech, (b) SMERSH, (c) spectral subtraction and (d) SMERSH + spectral subtraction.

the direct-to-reverberation ratio. For the results in this paper, $\kappa(k) = 1$. A method for blind estimation of $\kappa(k)$ is given in [9].

Finally, we estimate the clean speech signal by subtracting the late reverberation PSD from $|\tilde{X}(i, k)|$ according to

$$|\hat{S}(i, k)| = \max \left\{ |\tilde{X}(i, k)| - \sqrt{\lambda_l(i, k)}, |\tilde{X}(i, k)| G_{\min} \right\}, \quad (10)$$

and

$$\hat{S}(i, k) = |\hat{S}(i, k)| \frac{\tilde{X}(i, k)}{|\tilde{X}(i, k)|}. \quad (11)$$

The threshold constant $G_{\min}$ is introduced for robustness against overestimation errors and also controls the amount by which reverberation is attenuated. It is set here to $G_{\min} = -15$ dB. In the presence of noise, the late reverberant PSD estimate will be biased. When the noise is slowly time-varying the bias is equal to a fraction of the noise PSD [9]. Hence, the spectral subtraction will also suppress part of the noise. An unbiased estimator was proposed in [9].

III. SIMULATIONS AND RESULTS

A. Evaluation

Three objective measures were employed for evaluation of the processed speech:

(i) Segmental Signal to Reverberant Ratio (SRR) defined as

$$\text{SRR} = \frac{10}{I} \sum_{i=0}^{I-1} \log_{10} \left\{ \frac{\sum_{n=iN}^{iN+N-1} s_d^2(n)}{\sum_{n=iN}^{iN+N-1} (s_d(n) - \hat{s}(n))^2} \right\} \text{ dB}, \quad (12)$$

where $s_d(n)$ is the clean speech signal convolved with the direct path of the RIR, N is the frame length and I is the total number of frames considered.

(ii) Bark Spectral Distortion (BSD) defined as [16]

$$\text{BSD} = \frac{1}{I} \sum_{i=0}^{I-1} \frac{\sum_{b=0}^{N_b-1} (B_i(b) - \hat{B}_i(b))^2}{\sum_{b=0}^{N_b-1} B_i^2(b)}, \quad (13)$$

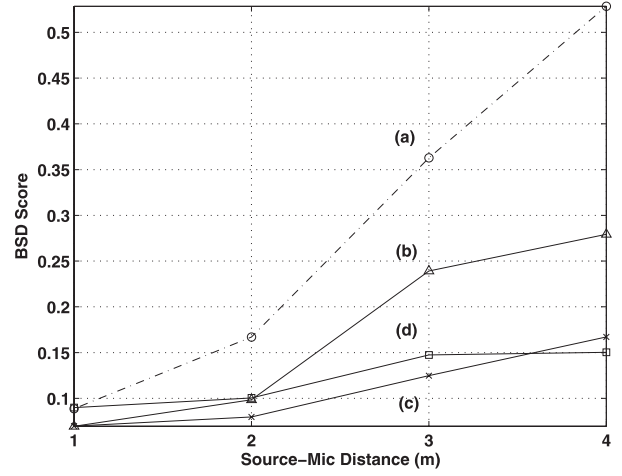


Fig. 2. BSD score for (a) reverberant speech, (b) SMERSH, (c) spectral subtraction and (d) SMERSH + spectral subtraction.

where $B_i(b)$ is the b th Bark subband from the i th speech frame and N_b is the total Bark subbands considered.

(iii) LP residual kurtosis measure calculated with

$$\text{Kurtosis} = \frac{1}{I} \sum_{i=0}^{I-1} \frac{E\{e_i^4(n)\}}{E^2\{e_i^2(n)\}} - 3, \quad (14)$$

where $e_i(n)$ is the i th frame of the prediction residual under consideration and $E\{\cdot\}$ is the expectation operator which is estimated from sample averages.

While metrics (i) and (ii) are conventional in speech processing, the Kurtosis measure is less so. It has been demonstrated to be a good indicator of reverberation [6], with its values decreasing as reverberation increases. This observation has been used to develop algorithms for reverberation reduction [6], [10] where the kurtosis is maximized adaptively. Results in [6], [10] indicate that this mainly reduces the early reflections, while the reverberation tail remains. Therefore, when used as a quantitative measure, it is a good indicator of the colouration in reverberant speech. A high correlation between this measure and colouration was verified in a recent study [17].

B. Simulation results

For the following simulations, we used measured RIRs from the MARDY database [18] for a linear array with $M = 8$ uniformly distributed microphones separated by 5 cm. The distance between the source and the center of the array varies in the range 1–4 m and the reverberation time is $T_{60} \approx 0.45$ s. The room is irregularly shaped with wall lengths 2.3, 9.3, 12 and 13.9 m and height 3 m. Anechoic speech samples were drawn from the APLAWD database [19], which contains five short sentences uttered by five male and five female talkers. These sentences were concatenated for each talker to produce one long utterance and all the results presented here are an average of the ten talkers. The samples also contain an EGG

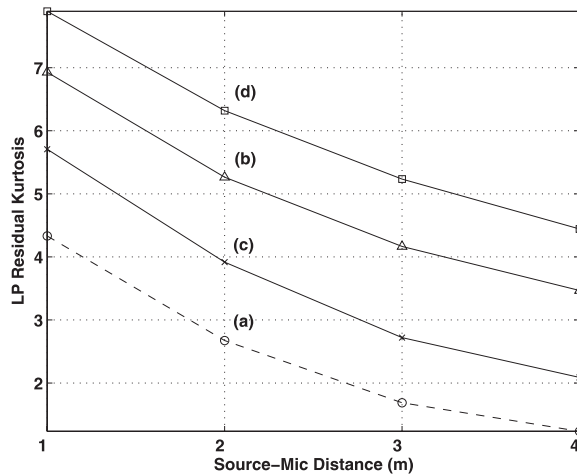


Fig. 3. LP residual kurtosis for (a) reverberant speech, (b) SMERSH, (c) spectral subtraction and (d) SMERSH + spectral subtraction.

signal which is used here for GCI identification using the HQT_x algorithm [20]. Reverberated samples were produced by convolution of the clean speech samples with the measured RIRs according to (1) and white Gaussian noise was added so that the SNR of the reverberant speech was 25 dB. The sampling frequency was set to $f_s = 8$ kHz.

Figures 1 and 2 show the simulation outcomes in terms of segmental SRR and BSD for (a) reverberant (unprocessed) speech, (b) speech processed with SMERSH, (c) speech processed with spectral subtraction (using only one microphone) and (d) the combination of SMERSH and spectral subtraction. It can be seen that spectral subtraction, generally, performs better than SMERSH when on its own. The combined system gains an additional 4 dB of segmental SRR compared with the spectral subtraction method. The two-stage approach and the spectral subtraction method appear to perform on a similar level in terms of BSD. Informal listening tests do not seem to correlate well with this result and indicate a greater reduction in reverberation using the combined approach. Although, some audible artefacts are present in the processed speech, both noise and reverberation are reduced. Illustrative listening examples can be found on: <http://www.commsp.ee.ic.ac.uk/~ndg/iscas08samples>.

Next, Fig. 3 shows the results in terms of the LP residual kurtosis. As expected, we see the opposite effect. SMERSH performs better than spectral subtraction in reducing the effect of the reflections and exhibits greater improvement over the reverberant speech. The combined system provides another level of improvement above either of the systems alone.

IV. CONCLUSIONS

We have presented a multimicrophone speech dereverberation algorithm where the processing considers two separate portions of the room impulse response. A spatiotemporal averaging method was used to primarily attenuate the early re-

verberation while spectral subtraction attenuates the remaining late reverberation. Simulation results using measured impulse responses and additive white Gaussian noise showed that the combined approach performs better or at least as well as each of the individual methods alone. Most importantly, the method does not require explicit knowledge of room impulse responses and was demonstrated to operate satisfactorily in the presence of measurement noise at SNR = 25 dB.

REFERENCES

- [1] P. A. Naylor and N. D. Gaubitch, "Speech Dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Eindhoven, The Netherlands, Sept. 2005.
- [2] Y. Huang, J. Benesty, and J. Chen, "A Blind Channel Identification-Based Two-Stage Approach to Separation and Dereverberation of Speech Signals in a Reverberant Environment," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 882–895, Sept. 2005.
- [3] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise Dereverberation Using Multichannel Linear Prediction," *IEEE Trans. Audio, Speech, Language Processing*, vol. 15, no. 2, pp. 430–440, Feb. 2007.
- [4] B. D. Radlović and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 728–737, Nov. 2000.
- [5] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8, no. 3, pp. 267–281, May 2000.
- [6] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, 2001, pp. 3701–3704.
- [7] H. Kuttruff, *Room Acoustics*, 4th ed. Taylor & Francis, Oct. 2000.
- [8] 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, pp. 607 – 610.
- [9] E. A. P. Habets, "Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007. [Online]. Available: <http://alexandria.tue.nl/extra2/200710970.pdf>
- [10] M. Wu and D. Wang, "A Two-Stage Algorithm for One-Microphone Reverberant Speech Enhancement," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 3, pp. 774–784, May 2006.
- [11] 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.
- [12] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPISA algorithm," *IEEE Trans. Audio, Speech, Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.
- [13] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*. Macmillan, 1993.
- [14] B. Yegnanarayana, J. M. Naik, and D. G. Childers, "Voice simulation: factors affecting quality and naturalness," in *Proc. Conf. of the Association for Computational Linguistics*, Stanford, California, USA, July 1984, pp. 530–533.
- [15] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. Springer-Verlag, Berlin, 2001.
- [16] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 819–829, June 1992.
- [17] E. A. P. Habets, N. D. Gaubitch, and P. A. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Las Vegas, USA, 2008.
- [18] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Paris, France, Sept. 2006.
- [19] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London, Tech. Rep., June 1987.
- [20] M. Huckvale, "Speech Filing System: Tools for Speech Research," [Online]. Available: <http://www.phon.ucl.ac.uk/resource/sfs/>, July 2003.