MULTI-SOURCE ESTIMATION CONSISTENCY FOR IMPROVED MULTIPLE DIRECTION-OF-ARRIVAL ESTIMATION

Sina Hafezi, Alastair H. Moore and Patrick A. Naylor

Department of Electrical and Electronic Engineering, Imperial College London, UK

{s.hafezi14, alastair.h.moore, p.naylor}@imperial.ac.uk

ABSTRACT

In Direction-of-Arrival (DOA) estimation for multiple sources, removal of noisy data points from a set of local DOA estimates increases the resulting estimation accuracy, especially when there are many sources and they have small angular separation. In this work, we propose a post-processing technique for the enhancement of DOA extraction from a set of local estimates using the consistency of these estimates within the time frame based on adaptive multi-source assumption. Simulations in a realistic reverberant environment with sensor noise and up to 5 sources demonstrate that the proposed technique outperforms the baseline and state-of-the-art approaches. In these tests the proposed technique had the worst average error of 9°, robustness of 5° to widely varying source separation and 3° to number of sources.

Index Terms— direction-of-arrival estimation, k-means clustering, estimation consistency, direct-path-dominance

1. INTRODUCTION

Multiple Direction of Arrival (DOA) estimation has been a popular though challenging field in acoustic signal processing. It has wide range of applications including source localization/tracking/separation, environment mapping, dereverberation and speech enhancement. In addition to reverberation and noise, extra factors such as number and separation of sources, their temporal activity, distance to microphone, loudness and motion make DOA estimation more challenging in multiple source scenarios compared to Single Source (SS) scenarios. In this work, we are interested in situations where we have multiple simultaneously active stationary sources with unknown separation, realistic reverberation and level of sensor noise. The number of sources is assumed to be known in prior.

For scenarios with speech sources, due to different timing, utterance, temporal and frequency characteristic of each speech source, there are often Time-Frequency (TF) bins with dominant SS in which a narrow band SS-based DOA estimator have high accuracy. Any SS-based DOA estimator from classes such as Steered Response Power (SRP)[1, 2, 3] or Intensity Vectors (IV)[4, 5, 6, 7] can be applied on each TF bin. There are two types of technique to extract multiple DOAs from the set of DOA estimates from all TF bins.

(1) Peak detection approaches in which the top peaks in the 2D histogram (azimuth vs elevation) of DOA estimates are directly [5, 6] or iteratively [8, 9] (by removing the contribution of the previously detected peak) extracted as the final DOAs. These approaches require spatial smoothing if the raw histogram is noisy, see for example Fig. 4(a). In order to obtain well-distinguished peaks and avoid merging two close peaks associated with two adjacent sources, the smoothing process needs to be adapted to the knowledge of the separation of sources [10]. It is also possible to have fewer peaks than the number of sources in a highly smoothed histogram [6]. Peak detection approaches are known to be prone to source separation [10] and therefore are not a reliable DOA extraction strategy for scenarios with unknown source separation with possibility of having adjacent sources.

(2) Clustering-based approaches such as K-means clustering [11, 10] classify the set of DOA estimates into clusters where their centroids represent the extracted DOAs. Although K-means clustering guarantees the detection of required number of centroids, it is shown in [10] to be highly prone to erroneous DOA estimates which are seen as noise in the histogram, see for example Fig. 5(b). It also suffers from the variation of solutions depending on centroids initialization. For a denoised histogram, clustering is more reliable and separation-independent than peak detection [10] and therefore is chosen as the DOA extraction strategy in this work.

In order to denoise the histogram of DOA estimates, a number of approaches such as Single Source Zone (SSZ) detection [8, 9] and Direct Path Dominance (DPD) test [12] have defined metrics which selects DOA estimates from certain TF bins as more reliable estimates. The state-of-the-art DPD method identifies SS dominant TF regions using the covariance matrix of the observed signals over a window centred on a TF bin. It is shown to have poor accuracy for low angular separation of $\leq 15^\circ$ [10] in unknown separation scenarios.

Our recently proposed technique, Estimation Consistency

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465

(EC) [10], uses the coefficient of variation (CV) [13, 14] to estimate the consistency of DOA estimates within a frame based on SS assumption, which is used to identify the TF bins with dominant SS. As shown in [10], the SS assumption in a time frame is rarely a good model in case of 4 or more simultaneously active sources. Therefore we propose an extension of our EC technique that is based on adaptive multiple source assumption in a time frame.

This paper is structured as follow: Section 2 briefly reviews the state-of-the-art DPD test. Section 3 presents our proposed multi-source technique and the SS-based EC technique which can be seen as a special case of the proposed technique. Finally in Section 4 we compare our proposed technique with the state-of-the-art and the baseline techniques.

2. DIRECT PATH DOMINANCE TEST

The DPD test [12] identifies the TF regions with significant contribution from the direct-path of a single source. It is based on the spatial covariance matrix $R_a(\tau, k)$ of the observed signals in a window on TF domain centred on the bin at time frame τ and frequency k. The TF bins are selected using

$$\Upsilon_{\mathrm{DPDtest}} = \{(\tau, k) : \mathrm{erank}\left(R_a(\tau, k)\right) = 1\}, \quad (1)$$

and

$$\mathrm{erank}\left(R_a(\tau, k)\right) = 1 \text{ if } \frac{\sigma_1(\tau, k)}{\sigma_2(\tau, k)} > threshold, \quad (2)$$

where $\mathrm{erank}(.)$ is the effective rank, $\sigma_1(\tau, k)$ and $\sigma_2(\tau, k)$ are respectively the largest and the second largest singular values of $R_a(\tau, k)$. Further information regarding the employment of DPD test on intensity-based DOA estimators can be found in [15].

3. PROPOSED METHOD

As shown in [10], the occurrence of the SS scenario in a time frame reduces as the number of simultaneously active sources increases. Instead of a SS assumption within a time frame, as in EC, or a TF region, as in DPD, we propose an adaptive Multi-Source Estimation Consistency (MSEC) model within a time frame.

Consider a time frame with unknown number of simultaneously active sources. We expect multiple concentrations of DOA estimates within the time frame if multiple sources are individually dominant at different frequencies. In such scenario, adaptive K-means clustering can be used to identify multiple concentrations of DOA estimates within a time frame.

3.1. Adaptive K-means clustering

Our adaptive K-means clustering uses Akaike Information Criterion (AIC) measure to find the best number of clusters. AIC trades-off distortion against the model complexity and is calculated as [16]

$$AIC(K) = -2L(K) + 2q(K), \quad (3)$$

where $-L(K)$, the negative maximum log-likelihood of the data for K clusters, represents a measure of distortion and $q(K)$, the number of parameters of the model with K clusters, represents a measure of model complexity.

Having performed K-means for $K = \{1, \ldots, K_{max}\}$ on a set of data with random initializations, the best number of clusters, K_c, is chosen as [17]

$$K_c = \arg\min_K \left[\mathrm{RSS}(K) + 2JK\right], \quad (4)$$

where $\mathrm{RSS}(.)$ is the residual sum of squared (sum of squared distance of each member to its cluster centroid) and J denotes the number of dimensions of centroid which leads to JK parameters for K clusters.

In order to have strong concentrations of DOA estimates for the purpose of robust clustering, adaptive K-means clustering at frame τ is performed on data set $U(\tau)$ including all DOA estimates from frame τ to $\tau - T$, which is defined as

$$U(\tau) = \{\hat{\mathbf{u}}(t, k) : \forall k,\, t \in \{\tau, \tau - 1, \ldots, \tau - T\}\}, \quad (5)$$

where $\hat{\mathbf{u}}(\tau, k)$ is the estimated DOA unit vector.

Having performed adaptive K-means clustering on data set $U(\tau)$, we obtain the optimum number of clusters $K_c(\tau)$, the clusters $\{S_i(\tau)\}_{i=1}^{K_C(\tau)}$ and the centroids unit vector $\{\hat{\mathbf{c}}_i(\tau)\}_{i=1}^{K_C(\tau)}$ where i is the cluster index. As a representative of the spread of DOA estimates within each cluster, the average member-to-centroid angular distance $D_i(\tau)$ is calculated for each cluster as

$$D_i(\tau) = \frac{1}{|S_i(\tau)|} \sum_{k \in S_i(\tau)} \angle(\hat{\mathbf{u}}(\tau, k), \hat{\mathbf{c}}_i(\tau)), \quad (6)$$

where $\angle(.)$ denotes the angle (in radians) between two vectors.

3.2. MSEC weighting

For each DOA estimate, the cluster weight, which represents the normalized measure of concentration in its associated cluster, is

$$\psi_{MS}(\tau, k) = 1 - \frac{D_i(\tau)}{\pi},\ k \in S_i(\tau), \quad (7)$$

and the member weight, which represents the normalized measure of closeness to its associated centroid, is

$$\lambda_{MS}(\tau, k) = 1 - \frac{\angle(\hat{\mathbf{u}}(\tau, k), \hat{\mathbf{c}}_i(\tau))}{\pi},\ k \in S_i(\tau). \quad (8)$$

The MSEC weight in the TF domain is then formed as

$$w_{MS}(\tau, k) = \psi_{MS}(\tau, k)\lambda_{MS}(\tau, k). \quad (9)$$

The EC weights [10], which is based on SS assumption within a time frame, can be calculated as a special case of MSEC with $K_{max} = 1$ and $T = 0$.

Figure 1 shows the spatial distribution of DOA estimates and their clusters for EC and MSEC. Unlike EC, where centroid in a time frame can lay far from sources due to the presence of multiple clusters as shown in Fig. 1(c), we have multiple centroids estimated close to sources in MSEC. We can also see the positive effect of including estimates from previous time frames for robust clustering as the clusters become more distinguished in Fig. 1(a) than Fig. 1(b).

Figure 2 shows the square root of MSEC (a) and EC (b) weights versus the normalized accuracy (to the nearest true DOA) for the DOA estimates from the same time frame used in Fig. 1. The positive effect of clustering can be seen as MSEC is significantly more correlated to accuracy compared to EC for time frames with simultaneously active sources.

3.3. DOA selection for the spatial histogram

Using EC or MSEC strategy, we select the DOAs associated with the TF bins with the top $M\%$ strongest weights to form the spatial histogram. Note that the weights are only calculated to rank the DOA estimates and not to weight them in the spatial histogram. In order to remove the highly isolated outlier DOAs in EC and MSEC spatial histograms, we remove the DOA estimates whose average angular distance to

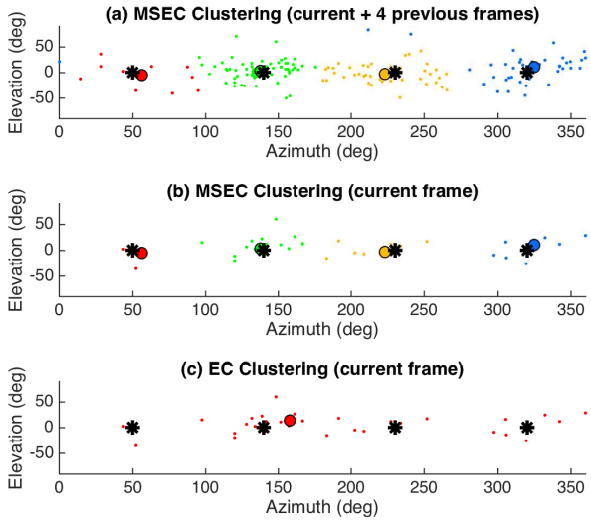

Fig. 1. DOA estimates (dot) distribution in space and their associated cluster (colour) and centroid (filled circle). (a) MSEC on all estimates from frame τ to $\tau - 4$. MSEC (b) and EC (c) on estimates from frame τ. True DOAs marked as black asterisk. 4 sources with 90° separation and $T_{60} = 0.4$ s.

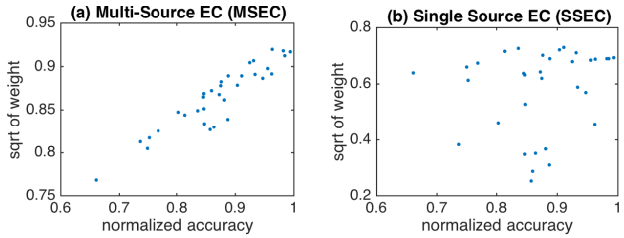

Fig. 2. Square root of MSEC (a) and EC (b) weights versus normalized accuracy for DOA estimates used in Fig 1.

L-nearest neighbours are less than η degrees.

4. EVALUATION

The method of Pseudo-intensity Vectors (PIV) [4] was used as our DOA estimator for Spherical Microphone Arrays (SMA) [18] as it gives a noisy spatial histogram. The Acoustic Impulse Responses (AIRs) of a 32-element rigid spherical microphone array were simulated using Spherical Microphone arrays Impulse Response Generator (SMIRgen) [19] based on Allen & Berkley's image method [20]. The array with radius 4.2 cm is placed at $(2.54,\ 4.48,\ 1.45)$ m in a $5 \times 6 \times 4$ m shoebox room with $T_{60} = 0.4$ s. N_s sources are placed on the same horizontal plane as SMA with azimuth separation $\triangle\phi$ with 1 m distance to the centre of SMA. Random DOAs were used for 100 Monte Carlo trials per each case of N_s and $\triangle\phi$. We employed anechoic speech sources randomly selected for each trial from the APLAWD database [21]. The active level of each speech source according to ITU-T P.56 [22] is set to be equal across all trials. Spatio-temporally white Gaussian noise is added to the microphone signals to produce a signal to incoherent noise ratio of 25 dB. A sampling frequency of 8 kHz was used with frame length of 4 ms and 50% overlapping of time frames.

In order to avoid any ambiguity due to data association uncertainty in our results, best case data association was used to obtain the mean estimation error where the error (in degrees) between a true DOA unit vector $\hat{\mathbf{u}}_o$ and an estimated DOA unit vector $\hat{\mathbf{u}}_s$ is

$$\varepsilon_{\mathbf{u_o},\mathbf{u_s}} = \cos^{-1}\left(\mathbf{u}_o^T\mathbf{u}_s\right). \quad (10)$$

In DPD, R_a in (1) and (2) was calculated using the spherical harmonic transformed signals with TF domain window size of 7 time frames and 5 frequency bins. The $threshold$ in (2) was empirically set to 6, which is also the recommended value in [12]. In MSEC, we had $K_{max} = 4$ and $T = 4$ frames. For EC and MSEC, M was empirically set to 10% with $L = 8$, $\eta = 5^\circ$ for outlier removal. For DOA extraction, we used one-run K-means clustering with N_s clusters, 100 maximum iterations and random initial centroids.

Figure 3 shows the mean error as a function of N_s and $\triangle\phi$ for the proposed methods EC and MSEC, DPD, as the state-

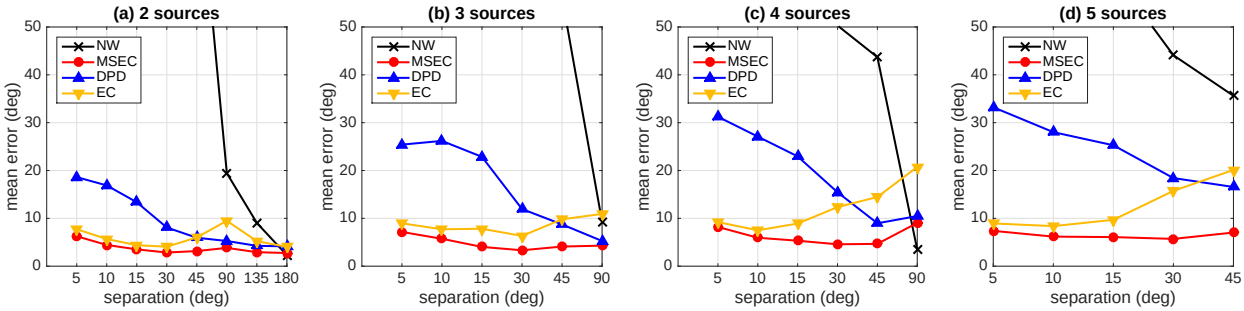

Fig. 3. Mean error of NW, MSEC, DPD and EC for varying N_s and $\triangle\phi$.

of-the-art and No Weighting (NW), as the baseline, in which the estimates from all TF bins are used in the histogram. N_s varies from 2 to 5 where in each case we have varying separation from 5° to the widest possible separation in our set. MSEC significantly outperforms all methods with the highest accuracy and robustness to number and separation of sources. NW, due to high noise in spatial histogram, fails almost in all cases except for separation of above 90°. DPD performs well with the accuracy of less than 10° for up to 4 sources with $\triangle\phi \geq 45^\circ$ while it loses accuracy for $N_s = 5$ or $\triangle\phi \leq 30^\circ$ due to the reason explained later. MSEC compared to EC, shows noticeably stronger robustness to separation as $\triangle\phi$ increases especially for 4 and 5 sources scenario. This is due to multi-source assumption within time frame in MSEC as the SS model in EC is rarely a good model within a frame for the case of many simultaneously active sources ($N_s = 4, 5$).

Figure 4 illustrates the 2D spatial histogram from side view angle for each method for a case of $N_s = 4$ and $\triangle\phi = 45^\circ$. Comparing NW with other methods, we can clearly observe the noise reduction in spatial histogram and its effect on the performance of clustering. We can also see the failure of EC in Fig 4 (c) as the estimated histogram is a bi-modal

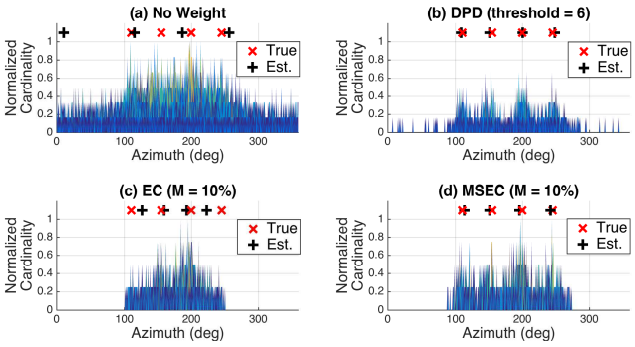

Fig. 4. Side view of spatial histogram for (a) NW, (b) DPD, (c) EC, (d) MSEC with $N_s = 4$, $\triangle\phi = 45^\circ$ and $T_{60} = 0.4\,\text{s}$. True DOAs and Estimated DOAs are marked as red ($\times$) and black ($+$) respectively.

distribution for 4 sources since the SS model in a time frame in EC mostly occurs for two of the sources. Unlike EC, DPD estimates a 4-modal distribution as shown in Fig 4 (b) since it is based on SS in TF region which is a more frequent assumption than SS in time frame. Figure 5 clearly shows the effect of outlier removal used for EC and MSEC, which results into clear histogram unlike DPD, which have some outliers. The outliers in DPD can result into wrong clustering in case of low separation as it leads to having a cluster for noise and a merged cluster for adjacent sources as seen in Fig. 5(b). This results into high estimation error as we can see in Fig. 3 for low separation.

5. CONCLUSIONS

We proposed a method for enhancement of accuracy and robustness to separation and number of sources of multiple DOA estimation that can be applied as a post-processing stage on any DOA estimator in TF domain. In case of unknown separation of sources, it outperforms the baseline and the state-of-the-art methods and provides strong accuracy of $\leq 9^\circ$ for up to 5 sources. It also shows the highest robustness of 5° to source separation and 3° to number of sources compared to other methods, which fail due to either low separation of sources or high number of sources.

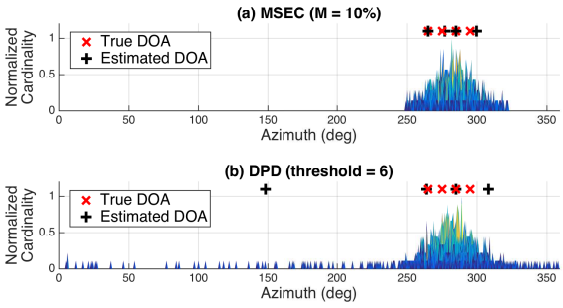

Fig. 5. Side view of spatial histogram for (a) MSEC, (b) DPD with $N_s = 4$, $\triangle\phi = 10^\circ$ and $T_{60} = 0.4\,\text{s}$.

6. REFERENCES

[1] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] B. Rafaely, "Plane-wave decomposition of the pressure on a sphere by spherical convolution," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2149–2157, Oct. 2004.

[3] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 361–371, Feb. 2011.

[4] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[5] S. Hafezi, A. H. Moore, and P. A. Naylor, "3D acoustic source localization in the spherical harmonic domain based on optimized grid search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Shanghai, China, March 2016.

[6] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, September 2016.

[7] D. Levin, E. A. P. Habets, and S. Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1800–1811, 2010.

[8] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker doa estimation in a circular microphone array based on matching pursuit," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Bucharest, Romania, August 2012.

[9] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3d localization of multiple sound sources with intensity vector estimates in single source zones," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, September 2015.

[10] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization using estimation consistency in the time-frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017.

[11] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Nice, France, July 2014.

[12] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[13] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2009, pp. 285–288.

[14] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, Nov. 2012.

[15] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[16] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[17] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, UK, 2008.

[18] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, Springer Topics in Signal Processing. Springer, Berlin Heidelberg, 2016.

[19] D. P. Jarrett, "Spherical Microphone array Impulse Response (SMIR) generator," http://www.ee.ic.ac.uk/sap/smirgen/.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[21] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Technical report, University College London, June 1987.

[22] ITU-T, "Objective measurement of active speech level," Dec. 2011.