SPEECH TEMPORAL DYNAMICS FUSION APPROACHES FOR NOISE-ROBUST REVERBERATION TIME ESTIMATION

Mohammed Senoussaoui

École de Technologie Supérieure
Département de Génie Logiciel et des TI
Montreal, Canada

*João F. Santos and Tiago H. Falk**

Institut National de la Recherche Scientifique
Centre ÉMT
Montreal, Canada

ABSTRACT

Reverberation and noise are known to be the two most important culprits for poor performance in far-field speech applications, such as automatic speech recognition. Recent research has suggested that reverberation-aware speech enhancement (or speech technologies, in general) could be used to improve performance. However, recent results also show existing blind room acoustics characterization algorithms are not robust under ambient noise and there is still room for improvement under such settings. In this paper, several fusion approaches are proposed for noise-robust reverberation time estimation. More specifically, feature- and score-level fusion of short- and long-term speech temporal dynamics features are proposed. With noise-aware feature-level fusion, gains of up to 15.4% could be seen in root mean square error. Score-level fusion, in turn, showed further improvements of up to 9.8%. Relative to a recently-proposed noise-robust benchmark algorithm, improvements of 30% could be seen, thus showing the advantages of speech temporal dynamics fusion approaches for noise-robust reverberation time estimation.

Index Terms— Reverberation time, speech enhancement, modulation spectrum, room acoustics, hands-free communications.

1. INTRODUCTION

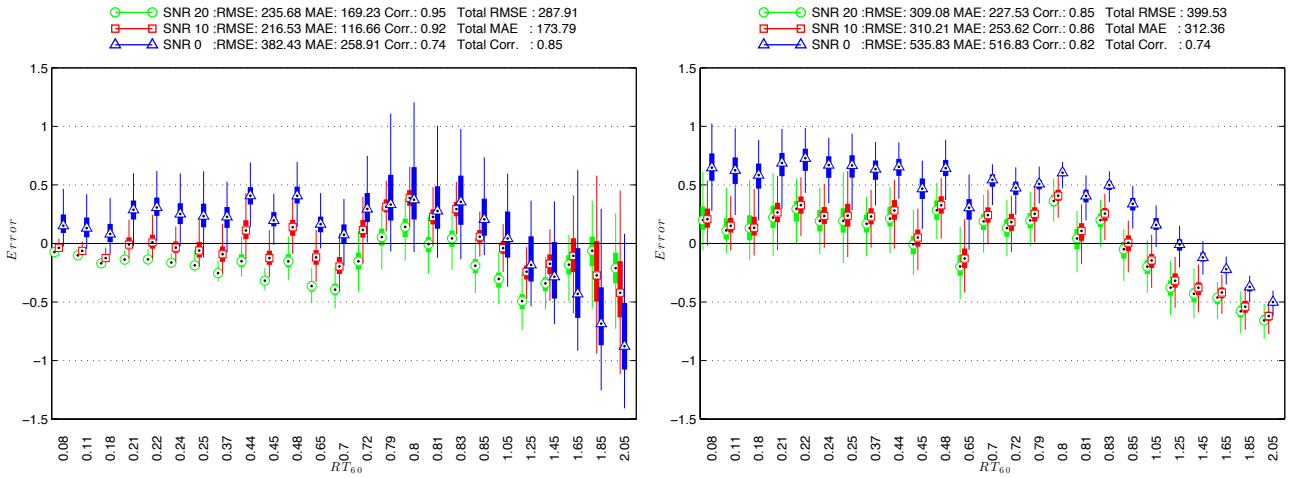
Speech technologies have left the research laboratory and are increasingly making their way into homes and offices. Today, a multitude of innovative voice-driven applications have emerged, transforming the way we interact with digital services and information. Automatic speech recognition (ASR), for example, has opened doors for automatic meeting transcription, in-vehicle control, smart TV and smartphone interaction, voice-based searches, to name a few applications. These speech applications are known to perform reliably in quiet rooms with close-talking microphones, but severe performance degradation occurs in more practical scenarios involving noisy environments and far-field microphones (e.g., hands-free teleconferencing and voice-enabled TV interaction) [1]. This degradation occurs mainly due to ambient noise and room reverberation, thus current research has aimed at developing advanced speech enhancement algorithms and/or noise-robust speech systems. Recent results, however, showed that further improvements are still needed in order to achieve acceptable performance [2].

Recently, it has been shown that environment-aware speech technologies can lead to improved performance. In [3], for example, the clarity index was used to improve speech recognition results. In turn, Reverberation Time (a common measure of room reverberation level, commonly termed RT60 or simply T60, as it quantifies the required time for the sound energy to decay by 60 dB after the extinction of the sound source) was used to select optimal models for far-field speech recognition [4] and speaker identification [5]. Measuring room acoustic parameters from speech, however, is not trivial, particularly in noisy environments, as recently shown by the Acoustic Characterization of Environments Challenge [6] and in [7].

Several "blind" approaches to estimating room acoustics parameters from speech recordings have been proposed in the literature over the past few years. Classical approaches have relied on estimating the time constant of the signal decay using maximum-likelihood (ML) approaches [8] or characteristics of the distribution of decay rates [9]. Due to their sensitivity to noise, updated versions of these methods have been recently proposed. In [10], an efficient T60 estimator based on a ML approach is proposed, where a smoothed histogram of the ML-predicted T60 for each frame is used to increase robustness of predictions; however, this is used in conjunction with a non-smoothed histogram of the last estimates, in order to enable faster updates of the prediction when T60 changes. The authors propose four other algorithms that improve this approach in [11], where subband information is exploited via a weighted average of the upper subband reverberation time (RT) estimates. In these updated algorithms, the authors do not use the fast-tracking of time-varying RT as it reduces noise robustness of the method. In [12], a SNR-dependent selection of time-frequency bins with higher likelihood of speech presence is proposed. Alternately, data-driven methods have been proposed where multiple features are extracted from the speech signal and mapped to different acoustic parameters (e.g., T60, C50) [13]. Short- and long-term speech temporal dynamics, in turn, were also shown to correlate with T60 [14] and reverberant speech quality [15]. Comparative analysis between different estimators has suggested that existing tools are sensitive to signal-to-noise ratio (SNR) levels [7], as well as to room reverberation levels [16].

In this paper, we propose different fusion strategies to improve T60 estimation in noisy environments. We build on the work of [14] and explore the fusion of short- and long-term speech temporal dynamics features, using both SNR-dependent and independent approaches. Experimental results show the proposed fusion method significantly improving blind room acoustics characterization performance in noisy scenarios and outperforming a noise-robust T60 estimation algorithm [12]. The remainder of this paper is organized as follow. Section 2 motivates the fusion of speech temporal dynamics information. The proposed fusion strategies then are described in Section 3. Sections 4 and 5 show the experimental setup and results, respectively. Lastly, conclusions are presented in Section 6.

*The authors acknowledge funding from NSERC, FRQNT, and the Nuance Foundation.



(a) (b)

Fig. 1. T60 estimation errors (seconds) obtained via a GLM mapping of four (a) short- and (b) long-term features.

2. FUSION OF SHORT- AND LONG-TERM SPEECH TEMPORAL DYNAMICS: MOTIVATION

The use of short- and long-term speech temporal dynamics was shown to be useful for T60 estimation in both clean and noisy environments [14, 7]. As in [14], the short-term dynamics are characterized by statistics computed from the delta coefficient of the zeroth order cepstral coefficient (a measure of the short-term log-spectral energy). Let $c_0(m)$ denote the zeroth order cepstral coefficient for frame m and $\Delta c_0(m)$ the zeroth order delta coefficient, thus

$$\Delta c_0(m) = \sum_{l=-L}^{L} l\, c_0(m+l), \tag{1}$$

where the normalization factor $\sum_{l=-L}^{L} l^2$ is omitted as it does not affect the results; in our simulations $L = 5$ is used. In order to capture short-term dynamics, sample statistics are computed from N Δc_0 samples (x_i). In particular, standard deviation (σ_Δ), skewness ($\mathcal{S}_\Delta$), kurtosis ($\mathcal{K}_\Delta$), and median absolute deviation ($\mathcal{D}_\Delta$) are computed according to

$$\sigma_\Delta = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}, \tag{2}$$

$$\mathcal{S}_\Delta = \frac{\sqrt{N}\sum_{i=1}^{N}(x_i-\bar{x})^3}{\left(\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^{3/2}}, \tag{3}$$

$$\mathcal{K}_\Delta = \frac{N\sum_{i=1}^{N}(x_i-\bar{x})^4}{\left(\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^{2}} - 3, \tag{4}$$

$$\mathcal{D}_\Delta = \text{median}_i(|x_i - \text{median}_j(x_j)|), \tag{5}$$

where $\bar{x}$ indicates the sample average of x_i. These four features are used as correlates of short-term dynamics.

Long-term temporal dynamics, in turn, are quantified via the modulation spectral representation described in [14, 15] and computed using the publicly available SRMRToolbox[1]. Here, only a brief description of the signal processing steps required is given; the interested reader is referred to [14, 15] for more details. First, the speech signal is filtered by a bank of 23 critical-band gammatone filters. The temporal envelope of each gammatone filter output is then calculated based on the Hilbert transform. Temporal envelopes are multiplied by a 256 ms Hamming window with 32 ms shifts and the modulation spectrum is then computed via a discrete Fourier transform. Lastly, modulation frequency bins are grouped into 8 bands. The modulation energy for gammatone channel j, modulation channel k, and frame m is given by $\mathcal{E}_{j,k}(m)$, $j = 1,\ldots,23; k = 1,\ldots,8; m = 1,\ldots,M$, where M denotes the total number of frames for a particular speech signal.

From [14], it was shown that modulation energy concentrated around $k = 1$ corresponded mostly to speech components, whereas for $k = 5-8$ to reverberation. As such, a per-band speech-to-reverberation modulation energy ratio (SRMR_k) was proposed and given by:

$$\text{SRMR}_k = \frac{\sum_{j=1}^{23}\sum_{m=1}^{M}\varepsilon_{j,1}(m)}{\sum_{j=1}^{23}\sum_{m=1}^{M}\varepsilon_{j,k}(m)}, k = 5,6,7,8. \tag{6}$$

These features are used as correlates of long-term dynamics.

The plots in Fig. 1 (as well as from Fig. 4 and Fig. 10 in [14]) motivate the proposed fusion approaches. A generalized linear regression (GLM) model was used to map the four short-term (plot a) and four long-term (plot b) features described above into T60 estimates for three different SNR levels using simulated reverberant speech with T60 levels ranging from 0.08 to 2.05 seconds. As can be seen, the short-term features are able to estimate accurately

[1] https://github.com/MuSAELab/SRMRToolbox

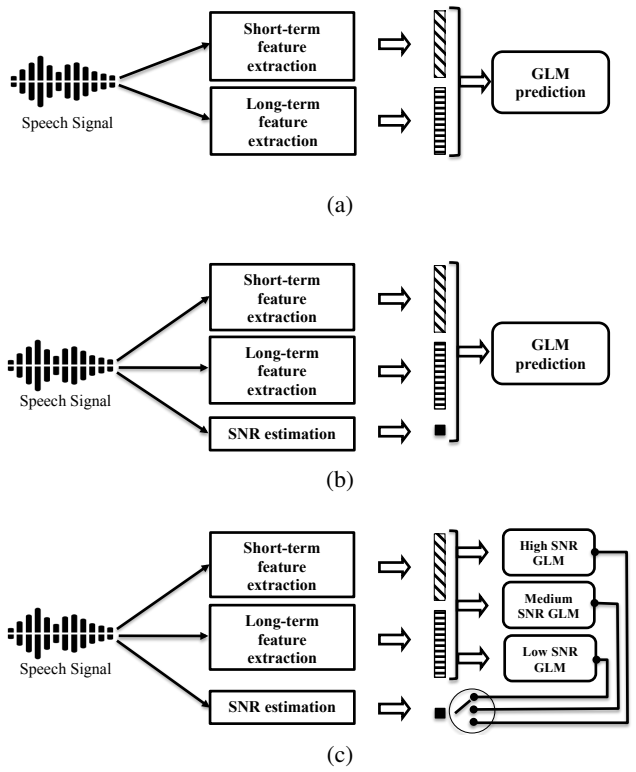


Fig. 2. Block diagram of proposed feature-level fusion strategies, including (a) SNR-independent, (b) SNR as a feature, and (c) SNR-aware fusion.

T60, particularly for lower reverberation levels. At approximately T60 $= 0.8$ s, however, estimation errors and estimation error variance increases, especially at lower SNR levels. For long-term features, on the other hand, an almost opposite behaviour is seen. For T60 > 0.8 s, estimation errors and error variances start decreasing, particularly for lower SNR levels. Such complementary behaviour suggest that the fusion of short and long-term features should improve overall T60 estimation accuracy over a wide range of reverberation levels. Moreover, estimator behaviour is shown to be sensitive to SNR levels, thus suggesting that an SNR-aware fusion strategy could further improve T60 estimation in noisy environments. These insights have motivated the proposed feature- and score-level fusion strategies described next.

3. PROPOSED FUSION STRATEGIES

3.1. Feature-level Fusion

Figure 2 (a-c) depicts the three feature-level fusion strategies explored here. The first (subplot a) depicts a simple, SNR-independent feature fusion scheme where the 4-D short-term feature vector is appended to the 4-D long-term feature vector to create a final 8-D vector. A GLM regressor based on a normal distribution and logarithmic link function is used to map the 8-D vector into a final T60 value. Next, two SNR-dependent approaches are explored. First, (subplot b), SNR is explored as an input feature to be fused with the short- and long-term features, thus resulting in a 9-D vector to be mapped via GLM regression to a final T60 value. Lastly, (subplot c) proposes an SNR-aware strategy where the SNR is used to shift between three GLM models, namely low (< 5dB), medium ($5 \leqslant$ SNR < 15 dB), and high ($\geqslant 15$ dB) SNR. Each model is trained on the 8-D fused short- and long-term features. In both cases, the SNR is estimated from the noisy speech signal using the noise analysis module of the International Telecommunication Union ITU-T P.563 single-sided speech quality measurement algorithm [17]. Estimation is performed by calculating the levels of speech and noise sections identified during voice activity detection; reliable SNR estimation accuracy was reported previously in [14]. These three T60 predictors (short, long, and fused) under the three fusion schemes (SNR-independent, SNR as a feature, and SNR-aware) are tested under different score-level fusion strategies, as described next.

3.2. Score-level Fusion

Here, three score-level fusion strategies are explored based on the outputs from the five different regressors described above (i.e short-term regressor, long-term regressor and three feature-level fusion regressors). The first is a simple averaging strategy, where two T60 estimates are averaged into a final value. In our experiments, all possible pair combinations were tested and only the one that achieved the highest accuracy is reported in Section 4. Second, a manual thresholding rule is explored based on the rule depicted by Fig. 3. More specifically, based on insights shown in Fig. 1, the following fusion rules are used: (1) if the estimated SNR is lower than 5 dB and the average estimated T60 (averaged from the predictors A and B, which could be any of the five described above) is (i) below 0.8 s, then the score from predictor A is used; otherwise (ii) the score from predictor B is used, and (2) if the estimated SNR is greater than 5 dB and the average T60 is (iii) below 0.8 s, then the scores from predictor B are used; otherwise (iv) the scores from predictor A are used. The advantage of these two score fusion strategies is that they do not rely on training of a separate fusion model. Lastly, a regression tree was used as an automated method to fuse the scores from two of the five potential T60 predictors. Here, a bootstrap aggregation (bagging) technique that is known to improve both the stability and predictive power of the trees was used [18, 19]. In order to keep the models simple and avoid overtraining, an ensemble of 25 trees was chosen empirically.

4. EXPERIMENTAL SETUP

In the experiments described herein, the TIMIT database was used. The original partitioning of training and test speakers (462 and 168, respectively) was kept to ensure unseen speakers during testing. For training of the GLMs and Regression Trees, the clean TIMIT training data was convolved with synthetic room impulse responses (RIR) created using the image method [20, 21] with T60 values ranging from $0.10 - 2.00$ s with 0.1 second increments. Next, a simulated speech shape noise generated from TIMIT train part was added to the reverberant signals at SNR levels of 0 dB, 10 dB, and 20 dB. The corrupted signals were then level-normalized to -26 dB overload (dBov) using the ITU-T P.56 voltmeter [22]. Next, the TIMIT clean test set was corrupted by a combination of synthetic and recorded RIRs. More specifically, the image method was used again to generate synthetic RIRs corresponding to T60 values ranging from $0.25 - 2.05$ s in increments of 0.2 s, thus resulting in values different from those available in the training set. Moreover, the recorded RIRs available in the Aachen Impulse Response (AIR) database [23, 24] were

Table 1. Performance comparison for proposed feature and score-level fusion strategies. Values in bold represent the fusion and scoring approaches that resulted in the lowest RMSE, MAE or ρ values.

	Feature fusion								
	SNR-independent			SNR as a feature			SNR-aware		
	RMSE	MAE	ρ	RMSE	MAE	ρ	RMSE	MAE	ρ
Short-term features	287.91	173.79	0.85	280.04	171.36	0.86	354.94	152.99	0.81
Long-term features	399.53	312.36	0.74	346.51	222.94	0.79	315.08	187.65	0.83
Feature Fusion	278.44	162.02	0.87	**262.57**	156.83	**0.88**	266.55	**115.92**	0.87
	Score fusion								
	Regression Tree			Manual Thresholding			Averaging		
	RMSE	MAE	ρ	RMSE	MAE	ρ	RMSE	MAE	ρ
Score Fusion	257.29	**109.32**	0.88	250.66	121.63	**0.89**	**240.49**	121.63	**0.89**

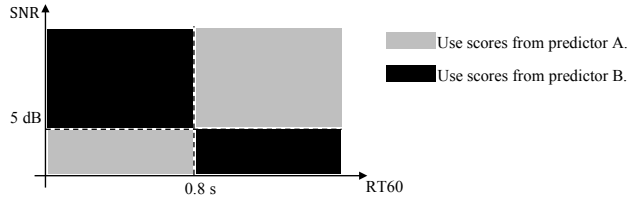


Fig. 3. Manual thresholding rule based on estimated SNR and T60. Predictors A and B can represent any of the three described in Section 3.1 or the individual regressors (i.e. short-term or long-term regressors).

also used to generate more realistic test signals with the following T60 values: $[0.08, 0.11, 0.18, 0.22, 0.24, 0.25, 0.44, 0.48, 0.72, 0.79, 0.80, 0.81, 0.83]$ (in seconds). Contrary to the training set, real metro station and restaurant noise taken from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [25] were added at SNR levels ranging from 0 dB to 20 dB. Similarly to what was done with the training set, the corrupted signals were level-normalized to -26 dB overload (dBov) using the ITU-T P.56 voltmeter.

In order to gauge the benefits of the proposed fusion schemes, three figures of merit are used. First, the Pearson correlation coefficient (ρ) computed between the estimated and true T60 values was used. Next, the root mean square error (RMSE) and the Median Absolute Error (MAE), both of them expressed in milliseconds, were used. As a benchmark, the noise-robust T60 estimation algorithm described in [12] was used.

5. EXPERIMENTAL RESULTS

The top part of Table 1 shows the experimental results obtained with the short- and long-term dynamics based predictors alone, as well as with feature fusion under the three different proposed setups. As can be seen, all feature fusion strategies resulted in improvements across all three figures of merit. The simple fusion, for example, resulted in a 6.7% decrease in MAE relative to the short-term feature based predictors. Moreover, the addition of the estimated SNR to the T60 predictors was also shown to be greatly advantageous. When using the estimated SNR as a feature, the RMSE dropped from 278.44 (simple fusion) to 262.57 ms, a 5.7% decrease. Relative to the short-term based predictor, a drop of almost 10% in MAE could be observed. With the SNR-aware strategy, in turn, these gains in MAE were of 33.3%, thus suggesting the advantages of SNR-aware fusion for T60 estimation. The benchmark algorithm, in turn, achieved the following performance metrics on the test set: $\rho = 0.81$, MAE$= 156.27$ ms, and RMSE$= 335.21$ ms. As such, gains of 25.8% could be achieved in MAE with the proposed SNR-aware system.

The lower part of Table 1, on the other hand, shows the experimental results obtained with the three score-based fusion schemes. For the simple averaging strategy, it was found that averaging the SNR-aware feature fusion method with the SNR as a feature fusion method resulted in the best performance. Overall, it achieved an RMSE of 240 ms, thus an additional 8.4% improvement over the feature fusion scheme with SNR as a feature. Relative to the benchmark, gains of 28.4% could be seen in the RMSE performance metric.

For the manual thresholding fusion scheme, it was found that combining the SNR-independent long-term dynamics based predictor with the SNR-aware feature fusion scheme resulted in the best results. As can be seen from the table, this fusion method was able to reduce RMSE to 250.66 ms, thus outperforming feature fusion (with SNR as a feature) by 4.5%. Both the manual thresholding and averaging fusion schemes achieved the same MAE and ρ values. Relative to the benchmark, gains of 22.1% could be seen in the MAE performance metric. Lastly, the regression tree ensemble method achieved the best results with a fusion of the SNR-aware long-term dynamics based system and the SNR-aware feature fusion system. Overall, this score fusion method achieved the lowest MAE of all tested methods, i.e., 109.32 ms. Relative to SNR-aware feature fusion and SNR-aware long term based predictors, this represented gains of 5.7% and 41.7%, respectively. Relative to the benchmark, gains of 30% could be seen in the MAE performance metric.

6. CONCLUSION

Previous work has shown the advantages of using short and long-term speech dynamics for blind room acoustics characterization, but sensitivity to ambient noise was reported. This paper has explored the complementarity of the two feature types for blind noise-robust reverberation time (T60) estimation. Several fusion and noise-aware strategies were explored. With feature fusion, it was shown that T60 estimate errors could be substantially reduced and the correlation between true and estimated T60 values increased. By incorporating estimated SNR values into the estimators, further gains could be achieved, as high as 28.5%. Lastly, score-fusion strategies were also proposed and additional gains (relative to feature fusion) as high as 8.4% could be seen. Relative to a widely-used noise-robust benchmark, a gain of 28.4% could be achieved in RMSE.

7. REFERENCES

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 114–126, 2012.

[2] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.

[3] P. P. Parada, D. Sharma, P. A. Naylor, and T. Van Waterschoot, "Reverberant speech recognition exploiting clarity index estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[4] J. Liu and G.-Z. Yang, "Robust speech recognition in reverberant environments by using an optimal synthetic room impulse response model," *Speech Communication*, vol. 67, pp. 65–77, 2015.

[5] A. R. Avila, M. O. S. Paja, F. J. Fraga, D. D. O'Shaughnessy, and T. H. Falk, "Improving the performance of far-field speaker verification using multi-condition training: the case of GMM-UBM and i-vector systems," in *INTERSPEECH*, 2014, pp. 1096–1100.

[6] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge—Corpus description and performance evaluation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*. IEEE, 2015, pp. 1–5.

[7] N. D. Gaubitch, H. W. Löllmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*. VDE, 2012, pp. 1–4.

[8] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[9] J. Y. Wen, E. A. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 329–332.

[10] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010, pp. 1–4.

[11] H. Löllmann, A. Brendel, P. Vary, and W. Kellermann, "Single-channel maximum-likelihood t60 estimation exploiting subband information," in *2015 ACE Challenge*. IEEE.

[12] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 161–165.

[13] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot, and P. Naylor, "A single-channel non-intrusive C50 estimator correlated with speech recognition performance," *IEEE/ACM Trans. Audio, Speech, and Language Processing*.

[14] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

[15] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1766–1774, 2010.

[16] F. Lim, P. A. Naylor, M. R. Thomas, and I. J. Tashev, "Acoustic blur kernel with sliding window for blind estimation of reverberation time," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*. IEEE, 2015, pp. 1–5.

[17] L. Malfait, J. Berger, and M. Kastner, "P. 563: The ITU-T Standard for Single-Ended Speech Quality Assessment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1924–1934, 2006.

[18] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996. [Online]. Available: http://dx.doi.org/10.1023/A:1018054314350

[19] ——, "Heuristics of instability and stabilization in model selection," *Ann. Statist.*, vol. 24, no. 6, pp. 2350–2383, 12 1996. [Online]. Available: http://dx.doi.org/10.1214/aos/1032181158

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979. [Online]. Available: http://dx.doi.org/10.1121/1.382599

[21] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1527–1529, 1986.

[22] *Objective Measurement of Active Speech Level*, 1993. [Online]. Available: http://www.itu.int/rec/T-REC-P.56-201112-I/en

[23] M. Jeub, M. Schafer, H. Kruger, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony?" in *International Congress on Acoustics (ICA)*, Agoust 2010.

[24] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *16th International Conference on Digital Signal Processing, 2009*, July 2009, pp. 1–5.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.