# BLIND ESTIMATORS FOR REVERBERATION TIME AND DIRECT-TO-REVERBERANT ENERGY RATIO USING SUBBAND SPEECH DECOMPOSITION

*Thiago de M. Prego*[1], *Amaro A. de Lima*[1], *Rafael Zambrano-López*[2], *and Sergio L. Netto*[2]

[1] Federal Center for Technological Education Celso Suckow da Fonseca, Nova Iguaçu, RJ, Brazil
[2] Federal University of Rio de Janeiro, Electrical Engineering Program-COPPE, RJ, Brazil
{thiago.prego, amaro.lima, rafael.lopez, sergioln}@smt.ufrj.br

## ABSTRACT

This paper describes algorithms for estimating two important features associated with the reverberation effect on speech signals: the reverberation time and direct-to-reverberant energy ratio. Both methods are referred to as blind algorithms in the sense that they are entirely based on the reverberant signal itself, not depending on the knowledge of the clean original signal. Proposed schemes use subband analysis to generate more and more reliable information, which is post-processed using basic statistical analysis to provide the desired estimate for each particular feature. Modifications on the original estimation algorithms are introduced to cope with lower SNRs. Performance of both algorithms is assessed under the ACE Challenge scope, which included a set of 288 speech signals for training and 4500 signals for final test. Results indicate the effectiveness of both techniques particularly in high-SNR situations.

***Index Terms***— blind estimation, reverberation time, direct-to-reverberant energy ratio, subband analysis

## 1. INTRODUCTION

Reverberation is an acoustical effect where the reflections of an audio signal on the room surfaces generate a series of attenuated-and-delayed copies which are perceived altogether as a single signal. Although a slight amount of reverberation can enhance the quality of speech signals, a long reverberation tail can severely affect speech intelligibility and/or reduce the perceived quality [1].

Among the main features associated to the reverberation effect, one may list the so-called reverberation time (RT, $T_{60}$) and the direct-to-reverberant energy ratio (DRR, $E_{dr}$). The $T_{60}$ is formally defined as the time interval required for a sound level to decay 60 dB after ceasing its original stimulus. Generally speaking, the RT quantifies the reverberation duration along time, whereas the DRR describes the reverberation effect in the space domain, providing insight on the relative positions of the sound source and receiver.

Most RT and DRR blind estimators found in the literature search for a persistent energy decay over time on the degraded signal, characterizing a so-called free decay region (FDR). In an FDR, one assumes that the sound stimulus has already finished and only the reverberation effect persists, allowing one to characterize it more clearly in such signal intervals.

The blind RT estimator presented in [2] performs a time-frequency decomposition of the speech signal using a sliding discrete Fourier transform (DFT). In each resulting subband, one searches for signal FDRs where an estimate for the RT and DRR features can be determined. In the end, the combined subband-FDR analysis generates more partial estimates for each feature which are subsequently sorted out by a basic statistical analysis, resulting in more reliable estimates.

The main contributions of this paper include adapting the subband RT estimation algorithm first presented in [2] to low-SNR conditions and introducing a similar subband estimation algorithm for the DRR feature. Performance assessment for both algorithms is then performed under the ACE Challenge framework including results for the algorithm development and evaluation stages [3].

To introduce the subband-FDR estimation algorithms for the RT and DRR reverberation features, this paper is organized as follows: In Section 2, the general subband-FDR scheme is introduced, whereas Section 3 discusses the modifications incorporated to the original algorithm making it suitable for low-SNR (below 30 dB, for instance) situations. Section 4 details the ACE Challenge, which considers the blind RT and DRR estimation for a large dataset of speech signals impaired by reverberation and additive background noise. Finally, Section 5 includes the results achieved by the proposed RT and DRR estimation algorithms under the ACE Challenge dataset conditions.

## 2. PROPOSED BLIND SUBBAND-FDR ESTIMATORS

The general framework for the proposed RT and DRR estimators is shown in Figure 1. In this diagram, each algorithm is implemented through five consecutive steps, which are detailed further below.

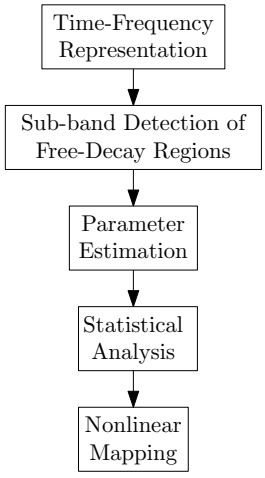


Figure 1: Block diagram of blind parameter (RT and DRR) estimators using subband decomposition.

978-1-4799-7450-4/15/$31.00 ©2015 IEEE

### 2.1. Time-Frequency Representation

In this initial stage, the reverberant speech signal, $s_r(n)$, is divided into $L$ frames using a length-$M$ window function $w(n)$, and a $K$-bin discrete Fourier transform (DFT), $\mathcal{F}\{\cdot\}$, is applied to each frame, generating the time-frequency representation $S_r(k,l)$ such that

$$S_r(k,l) = \mathcal{F}\{w(n)s_r(n)\}, \tag{1}$$

for $k = 0, 1, \ldots, (K-1)$, $l = 0, 1, \ldots, (L-1)$, and $n = l(M-V)$, $l(M-V)+1$, $\ldots$, $l(M-V)+M-1$, where $V$ is the number of overlapping samples of two consecutive frames.

In our spectral analysis, $K$ was chosen as the lowest power-of-two greater than or equal to $M$. As most of the speech energy lies within the analog frequency range $0 \leq f \leq 4$ kHz, we restrict all subsequent analyses to the values of $k$ such that $0 \leq \frac{F_s k}{K} \leq 4$ kHz, thus achieving more reliable feature estimates, where $F_s \geq 8$ kHz is the associated sampling frequency.

### 2.2. Subband FDR Detection

As mentioned in Section 1, the FDRs are characterized by a consistent energy drop in consecutive signal frames. In the proposed algorithm, this search is performed for all individual subbands, which tend to present a distinct energy pattern [4]. By defining the energy of the $k$th sub-band of the $l$th signal frame as

$$E(k,l) = |S_r(k,l)|^2, \tag{2}$$

the FDR search is performed across the frame index $l = 0, 1, \ldots, (L-1)$, for each frequency bin $k$. Extending Vieira's criterion [5, 6] to the transform domain, a subband-$k$ FDR is characterized by a decrease of $E(k,l)$ within a 500-ms interval along $l$. Using frames of $M = 0.05\,F_s$ samples with $V = M/4$ overlapping samples, this 500-ms interval translates into consecutive

$$L_{\lim} = \frac{0.500\,F_s}{M-V} \approx 13 \tag{3}$$

subband frames with decreasing energy. In the proposed algorithm, however, if no FDR satisfies this criterion in a given subband, this threshold number $L_{\lim}$ is iteratively reduced to at least 3 frames, guaranteeing a minimum amount of meaningful data for the subsequent stages of the algorithm [2].

For a sampling frequency of $F_s = 16$ kH as employed in the ACE Challenge, for instance, we have a frame length of $M = 0.05\,F_s = 800$ samples and, therefore, $K = 1024$ spectral bins.

### 2.3. Subband Feature Estimation

Standard RT-estimation algorithms search for the time interval required by some linear fitting of the energy decay function (EDF)

$$c(n) = 10\log_{10}\left(\frac{\sum_{\nu=n}^{N-1} h^2(\nu)}{\sum_{\nu=0}^{N-1} h^2(\nu)}\right) \text{ dB}, \tag{4}$$

for $n = 0, 1, \ldots, (N-1)$, to drop 60 dB [7, 8, 9]. The key aspect in such algorithms is choosing a proper time interval $n_1 \leq n \leq n_2$ to perform the linear EDF approximation. In general, $n_1$ is taken as such $c(n_1) = -5$ dB [10], whereas $n_2$ minimizes the mean-squared error (MSE) between the actual value of $c(n)$ and its linear fitting within $n_1 \leq n \leq n_2$ [8, 9]. In our RT estimator, we have adapted Schroeder's algorithm [7] to the frame-based subband EDF (SEDF) defined as

$$\bar{c}(k,l) = 10\log_{10}\left(\frac{\sum_{\lambda=l}^{\bar{L}-1} E(k,\lambda)}{\sum_{\lambda=0}^{\bar{L}-1} E(k,\lambda)}\right) \text{ dB}, \tag{5}$$

for $l = 0, 1, \ldots, (\bar{L}-1)$, where $\bar{L}$ is the number of frames within a given subband FDR. Therefore, the RT estimate is determined by the time interval required by an SEDF linear fitting to drop 60 dB, with the frame-index extremes chosen analogously as before [2].

For a length-$N$ RIR $h(n)$, the DRR is defined as [11, 12]

$$E_{dr} = \frac{\sum_{n=n_d-n_a}^{n_d+n_b} h^2(n)}{\sum_{n=n_d+n_b}^{N} h^2(n)}, \tag{6}$$

where $n_d$, $n_a$, and $n_b$ are the time indexes associated to the direct-sound, 1-ms, and 1.5-ms components, respectively. Within the $k$th spectral bin of a given FDR, the maximum value of $S(k,n)$ along $n$ is associated with the direct-sound component, and the $k$th subband DRR estimate for the $r$th FDR is given by

$$\hat{E}_{dr}(r,k) = \frac{\sum_{n=n_d(k)-n_a}^{n_d(k)+n_b} S^2(k,n)}{\sum_{n=n_d(k)+n_b}^{N_r} S^2(k,n)}, \tag{7}$$

with $n_a$ and $n_b$ as before. There could be more than one FDR for the same subband, making a total of $R_k$ FDRs for the whole speech signal.

The ACE challenge adopted a DRR definition based on [13], which is different although similar from the one described in (6):

$$\hat{E}_{dr}(r,k) = \frac{\sum_{n=n_d(k)-n_a}^{n_d(k)+n_b} S^2(k,n)}{\sum_{n=0}^{n_d(k)-n_a} S^2(k,n) + \sum_{n=n_d(k)+n_b}^{N_r} S^2(k,n)}, \tag{8}$$

where $n_a = n_b$ is the time index associated to an 8-ms length. In accordance to the ACE challenge's DRR definition, our approach was modified to use (8) instead of (7).

### 2.4. Statistical Analysis

The role of the statistical-analysis stage is to sort out all the subband partial estimates to generate a reliable candidate for the final RT and DRR estimates.

Assuming that a total of $R_k$ FDRs were found in the $k$th subband, each partial RT estimate can be denoted by $\hat{T}_{60}(r,k)$, for $r = 1, 2, \ldots, R_k$. In the proposed scheme, the estimate $\bar{T}_{60}$ is determined as the median value of all subband medians $\bar{T}_{60}(k)$, thus

avoiding biased/noisy extreme values generated in previous stages of the algorithm.

The DRR estimate follows the same rationale of the RT estimate. The $k$th subband DRR estimate, $\bar{E}_{dr}(k)$, is obtained from the average of all partial DRR estimates, $\hat{E}_{dr}(r,k)$, with FDRs associated to the subband in analysis. The proposed DRR estimate $\hat{E}_{dr}$ is determined as the average value of all subband DRR estimates $\bar{E}_{dr}(k)$ restricted to the frequency range $2.5 \leq f \leq 3$ kHz.

### 2.5. Parameter Mapping

The relationship between the sub-band ($\bar{T}_{60}$) and fullband ($\hat{T}_{60}$) RT estimates constitutes an open problem in the literature [14, 15, 16]. Our subband RT estimates, for instance, although highly correlated to the standard fullband $T_{60}$ metric, vary within a different dynamic range due to the median operator employed in its derivation. To compensate for this, a scale-adjusting linear-regression mapping is performed using the ACE Challenge development dataset (see Section 4 below) onto the feature estimates such that

$$\hat{T}_{60} = \alpha \bar{T}_{60} + \beta, \tag{9}$$

with $\alpha = 3.2$ and $\beta = -962$ ms for estimated SNR greater or equal than 5 dB and $\alpha = 4.1$ and $\beta = -1173$ ms for estimated SNR lower than 5 dB chosen in a system training stage.

For the DRR estimate, the linear mapping is of the form of (9) with $\alpha = 4.4$ and $\beta = 15.4$ dB for estimated SNR greater or equal than 5 dB and $\alpha = 2.7$ and $\beta = 12.8$ dB for estimated SNR lower than 5 dB.

## 3. ALGORITHM ADAPTATION FOR LOW SNR LEVELS

### 3.1. Denoising

The denoising approach adopted in this work requires an SNR estimation algorithm, which is an extremely simple technique. Assuming that the first 500 ms of the speech signal provided by ACE Challenge had only noise samples, the noise power $P_n$ and its standard deviation $\sigma_n$ were calculated over a few of the ACE development signals. Then, the remaining of each signal is split into frames of 5 ms, and each frame energy is compared to the threshold $(P_n + \sigma_n)$: if the frame energy is greater, the frame is labeled as speech corrupted by noise; otherwise, it is labeled only as noise. In order to enforce some time consistency, intervals lower than 50 ms of either noise or speech are considered mislabeled. By knowing the intervals of speech corrupted by noise and only noise, one can estimate their respective powers $P_{s+n}$ and $P_n$, and form an initial SNR estimation

$$\bar{\rho} = 10 \log_{10} \frac{P_{s+n} - P_n}{P_n}. \tag{10}$$

The final SNR estimate was obtained applying a linear mapping $\hat{\rho} = \alpha\bar{\rho} + \beta$, with $\alpha = 1.1$ and $\beta = -5.6$, which reached 96% correlation with only a 2.29 dB of root mean-squared error (RMSE) using the development data provided by the ACE Challenge staff.

The original RT and DRR estimation algorithms were drastically affected by the SNR inflicted in the speech signal, due to the fact that they were originally designed to work in a scenario of high SNRs. In order to adapt both algorithms to distinct SNR conditions, two different procedures were considered:

- In the high-SNR case ($\hat{\rho} \geq 5$ dB), the speech signal was preprocessed by a noise tracker [17] algorithm followed by a standard "direct-decision" speech estimator [18], which was capable of efficiently reducing the noise in such low-degradation scenario.
- In the low-SNR case ($\hat{\rho} < 5$ dB), a minimum MSE algorithm was employed to estimate the spectral amplitude of the signal [18, 19], which has shown to be effective in harsh conditions, followed by the same noise-tracker algorithm as before.

### 3.2. SNR-Based Linear Mapping

In the original RT estimation approach [2], devised for high SNR scenarios (above 30 dB), the linear mapping (9) considered $\alpha = 3.4$ and $\beta = -1165$ ms. The same linear mapping, however, was changed to as given in Subsection 2.5, to comply with the lower-SNR scenarios covered by the ACE Challenge conditions. A similar SNR-dependant mapping was also employed by the proposed DRR algorithm.

## 4. ACE CHALLENGE

The Acoustic Characterization of Environments (ACE) Challenge is a competition devised to evaluate state-of-the-art algorithms for blind acoustic $T_{60}$ and DRR estimation from speech signals.

A dataset specifically designed for the challenge was provided, including speech with durations between a few seconds and over a minute, from male and female talkers in different sized rooms and noise conditions (ambient, fan, and babble noise) for a single or multiple microphone(s). The dataset is divided into two databases:

- Development (Dev) database: composed by a set of 288 noisy reverberant speech files. It was also provided the ground truth values for the $T_{60}$ and DRR measurements.
- Evaluation (Eval) database: this database comprise a broader range of reverberation conditions. It is composed by a set of 4500 noisy reverberant speech files for each microphone configuration. Naturally, no ground truth is provided for this dataset.

The noisy reverberant speech files were constructed from anechoic speech convolved with the measured acoustic impulse responses (AIRs) obtained from a given room, with additive noise recorded in the same room conditions. Both Dev and Eval datasets were downsampled to a sample rate of $F_s = 16$ kHz and converted to 16-bit depth.

## 5. ESTIMATION RESULTS

The algorithms described in this paper were applied only to the single-microphone configuration of the ACE Challenge.

Figure 2 shows the results of the estimated parameters for the 288 signals of the Dev database, combining with the provided ground truth $T_{60}$ and DRR measurements for each SNR.

The estimated error of $T_{60}$ and DRR parameters along de Dev database is shown in Figures 3 and 4 respectively. From this data, one notices that the proposed algorithm is more effective when estimating the $T_{60}$ (which reach 90% correlation on high SNR) than the DRR measures.

The results of the ACE Challenge fullband $T_{60}$ estimation error and DRR percentage estimation error for each type of noise and SNR are shown in Figures 5 and 6, respectively. Comparing the results obtained for the development and evaluation datasets, one clearly notices a significant increase in the estimation error dynamic range for both features, what should be expected since the Eval dataset spans a wider range of reverberation scenarios. Such increase is more noticeable in the DRR estimates. For the RT estimator, however, one can notice an excellent performance of the proposed estimator, particularly for higher SNR values, indicating a good generalization property for the proposed algorithm.

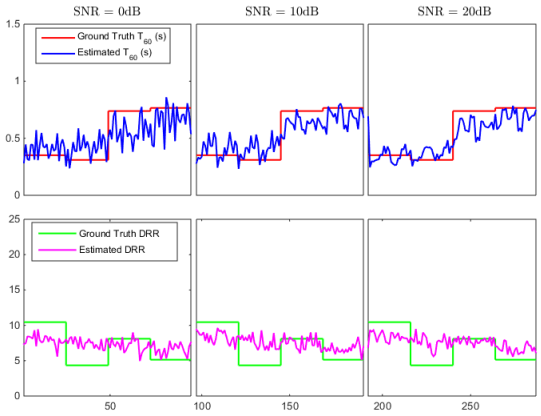


Figure 2: Estimated $T_{60}$ and DRR combined with ground truth parameters on Dev database.

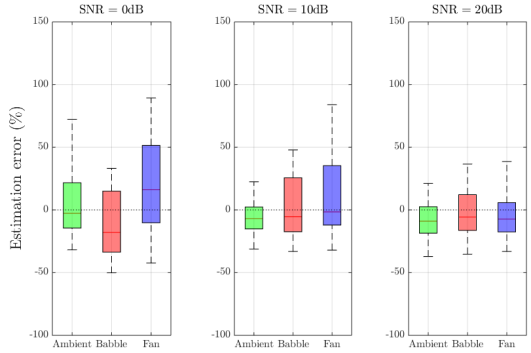


Figure 3: $T_{60}$ percentage estimation error by types of noise and SNR on Dev database.

## 6. CONCLUSION

This paper described blind algorithms for estimating the reverberation time (RT) and direct-to-reverberant energy ratio (DRR) of speech signals. Both algorithms employ a subband analysis and search for free-decay regions in each spectral bin to generate more partial estimates for a subsequent data-sorting procedure. Performances of the two algorithms were assessed under the ACE challenge conditions. Results indicated a reasonable DRR estimate and a quite successful RT estimation, particularly for higher SNR levels.

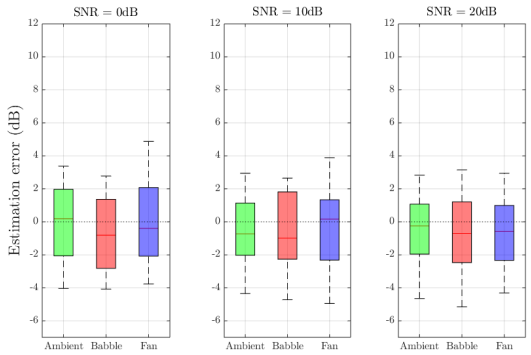


Figure 4: DRR estimation error by types of noise and SNR on Dev database.

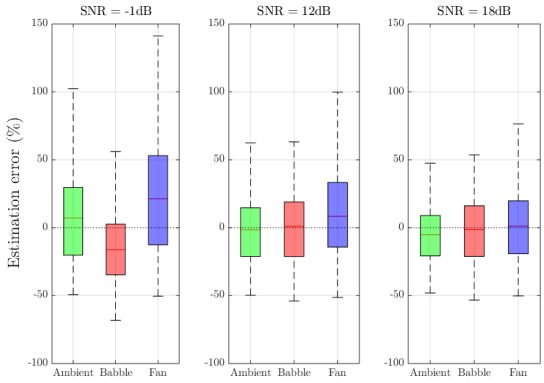


Figure 5: $T_{60}$ percentage estimation error by types of noise and SNR on Eval database.

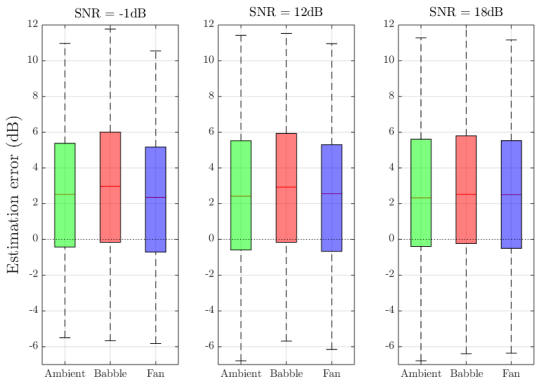


Figure 6: DRR estimation error by types of noise and SNR on Eval database.

## 7. REFERENCES


[1] R. Appel and J. Beerends, "On the quality of hearing ones own voice," *J. Audio Engineering Soc.*, vol. 50, no. 4, pp. 237-248, 2002.

[2] T. de M. Prego, A. A. de Lima, S. L. Netto, B. Lee, A. Said, R. W. Schafer, and T. Kalker, "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *J. Acoustic. Soc. Am.*, vol. 131, no. 4, pp. 2811–2816, 2012.

[3] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE Challenge - Corpus description and performance evaluation," *Proc. IEEE Workshop Applic. Sgnl. Process. Audio and Acoustics*, New Paltz, USA, Oct. 2015.

[4] L. L. Beranek, "Concert hall acoustics – 1992," *J. Acoustic. Soc. Am.*, vol. 92, no. 1, pp. 1–39, July 1992.

[5] J. Vieira, "Automatic estimation of reverberation time," *Proc. Conv. Audio Engineering Soc.*, Berlin, Germany, pp. 1–7, May 2004.

[6] J. Vieira, "Estimation of reverberation time without test signals," *Proc. Conv. Audio Engineering Soc.*, Barcelona, Spain, pp. 1–7, May 2005.

[7] M. R. Schroeder, "New method of measuring reverberation time," *J. Audio Engineering Soc.*, vol. 37, no. 3, pp. 409–412, 1965.

[8] N. Xiang, "Evaluation of reverberation times using a nonlinear regression approach," *J. Acoustic. Soc. Am.*, vol. 98, pp. 2112–2121, Oct. 1995.

[9] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *Proc. Conv. Audio Engineering Soc.*, Amsterdam, Netherlands, pp. 867–878, May 2001.

[10] ISO Rec. 3382, *Measurement of the reverberation time of rooms with reference to other acoustical parameters*, 1997.

[11] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J. Acoustic. Soc. Am.*, vol. 111, pp. 1832–1846, 2002.

[12] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Acoustic. Soc. Am.*, vol. 112, pp. 2110–2117, 2002.

[13] S. Mosayyebpour, H. Sheikhzadeh, T. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based inverse filtering of the room impulse response," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1617–1632, 2012.

[14] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Las Vegas, USA, pp. 329–332, Apr. 2008.

[15] T. H. Falk and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," *Proc. IEEE Int. Workshop Acoustic Echo and Noise Control*, Seattle, USA, pp. 1–4, Sept. 2008.

[16] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," *Proc. Workshop Circuits, Systems and Signal Processing*, Veldhoven, Netherlands, pp. 250–254, Nov. 2004.

[17] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1112–1123, 2008.

[18] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109-1121, 1984.

[19] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 2013. Acessed in May 2013, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.