ENVIRONMENT-INVARIANT COMPENSATION FOR REVERBERATION USING LINEAR POST-FILTERING FOR MINIMUM DISTORTION

Kshitiz Kumar[1] *and Richard M. Stern*[1,2]

Department of Electrical and Computer Engineering[1]
Language Technologies Institute[2]
Carnegie Mellon University, Pittsburgh, PA 15213
Email: {kshitizk,rms}@ece.cmu.edu

ABSTRACT

Speaker identification systems work quite well in controlled environments but their performance degrades severely in the presence of the reverberation that is frequently encountered in realistic acoustical environments. In this paper we develop an algorithm to make speaker identification systems more robust to reverberation by passing sequences of cepstral features through a short FIR filter. The coefficients of the filter are chosen to minimize the mean square differences between compensated features in the training and testing environments. Surprisingly, the resulting filter coefficients are relatively invariant to the actual nature of the reverberation. The use of the post-filtering approach is shown to improve speaker identification accuracy, especially when reverberation times are relatively long.

***Index Terms*—** Speaker Recognition, Deconvolution, Least Mean Square Methods, Wiener Filtering

1. INTRODUCTION

Speaker recognition or speaker identification (SID) is the technology that attempts to identify a subject on the basis of his or her speech [1]. Some of the important applications of the SID include speaker authentication, security, and verification.

Current state-of-the-art SID systems perform very well in controlled environments where speech samples collected for the identification task are reasonably clean, but real life environments are far less controlled. SID accuracy can deteriorate significantly in the presence of noise, interference, and reverberation. Robustness to noise and channel mismatch has been studied in [2][3], which include the application of algorithms such as cepstral substraction, feature warping, and feature transformation to SID. In this paper, we study the robustness of SID to reverberation. The issue of reverberation has been studied in terms of multi-microphone array processing in [4]. Compensation methods based on score fusion have been developed in [3]. Compensation based on multi-style training for different reverberation conditions is described in [5].

Many of the previous approaches for reverberation compensation either require multiple microphones or training data from different environments. The solution for reverberation compensation is usually local and no guarantees are made about performance across different conditions. Our approach for reverberation compensation is based on a single microphone, and we attempt to develop a solution which is global in scope that provides improvement across a wide range of reverberant environments.

The rest of the paper is organized as follows. We begin with a mathematical characterization of the effects of reverberation in Sec. 2. In Sec. 2.2 we describe and optimize our algorithm for reverberation compensation based on post-filtering of cepstral sequences. Experimental results are described in Secs. 3.1 and 3.2, and we discuss the underlying assumptions in our algorithm in Sec. 4. Section 5 summarizes the findings of this study.

2. SPEAKER IDENTIFICATION IN REVERBERANT ENVIRONMENTS

In this section, we describe some of the effects of reverberation. We propose a representation which relates reverberated speech to clean speech. This representation leads to a distortion measure which characterizes the mismatch between reverberated and clean speech. Next, we propose a solution that is based on abstractly passing the reverberated and clean speech through a linear filter and determining the coefficients that minimize the mean squared distortion. We solve this optimization in section 2.2, and we show that under certain assumptions our solution is unique, optimal, and invariant to the actual reverberation, so that the same solution works across different conditions.

2.1. Mathematical Representation of Reverberation

This section provides a representation of reverberated speech in terms of the corresponding clean speech. The SID system works in conventional fashion, by extracting features from the signal and determining which of a set of trained models provides the best match to an ensemble of incoming features. In order to maintain high SID accuracy, it is desirable that the features derived from reverberated speech closely match features derived from the corresponding clean speech. Let $x[n]$ represent the *cepstral features* of a speech waveform (and not the original waveform itself), and let $y_u[n]$ represent the corresponding cepstral features after undergoing room reverberation. Because reverberation can be thought of as the convolution of the input speech with the effective impulse response of a room, there would be a constant difference between $x[n]$ and $y_u[n]$ if these features represented long-term cepstra of the entire waveform. When $x[n]$ and $y_u[n]$ are cepstral coefficients of brief segments of speech (as in short-time Fourier analysis), there is an interaction between the speech and the analysis window and the difference between $x[n]$ and $y_u[n]$ is no longer constant. For simplicity, we propose that the reverberated cepstral features $y_u[n]$ can be represented as the convolution between the input cepstra $x[n]$ and the cepstral coefficients

$h[n]$ representing the effects of the room:

$$y_u[n] = x[n] + \sum_{i=1}^{N_h} h[i]x[n-i] \tag{1}$$

Referring to Fig. 1, $\boldsymbol{x} = x[n]|_{n=1}^{M}$ represents clean speech features and $\boldsymbol{y_u} = y_u[n]|_{n=1}^{M}$ represents the corresponding reverberated features. The subscript u in $\boldsymbol{y_u}$ indicates *uncompensated* speech in the testing environment. Thus, the assumption in (1) implies a linear filter in the cepstral feature domain with filter taps being $\boldsymbol{h} = [1\, h_1 \cdots h_{N_h}]$. Note that in this representation $h[0] = 1$. As a result of reverberation, system training will be performed on the features of clean speech $\boldsymbol{x}$ but testing will use reverberated features $\boldsymbol{y_u}$. We define the instantaneous *uncompensated distortion* $d_u[n]$ to be

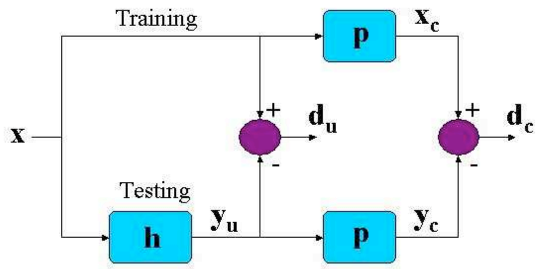

Fig. 1. Block diagram representing the model of reverberation and compensation.

$$d_u[n] = x[n] - y_u[n] = x[n] - h[n] * x[n] = \sum_{i=1}^{N_h} h[i]x[n-i] \tag{2}$$

(The second equality is valid because $h[0] = 1$.)

We compensate for the effects of reverberation by imposing a finite-impulse response LTI filter on the observed features ($\boldsymbol{p}$ in Fig. 1). We refer to the outputs of these features as *compensated*, and we use the notations $\boldsymbol{x_c}$ and $\boldsymbol{y_c}$ to indicate features representing compensated clean speech and reverberated speech, respectively. We define the instantaneous *compensated distortion* $d_c[n]$ to be the difference between $x_c[n]$ and $y_c[n]$:

$$\begin{aligned} d_c[n] &= x_c[n] - y_c[n] = p[n] * x[n] - p[n] * h[n] * x[n] \\ &= p[n] * d_u[n] = \sum_{j=0}^{N_p-1} p[j]d_u[n-j] \end{aligned} \tag{3}$$

where N_p is the number of taps in the $\boldsymbol{p}$ filter. We seek to obtain the optimal $\boldsymbol{p}$ filter which, when applied to both $\boldsymbol{x}$ and $\boldsymbol{y_u}$, minimizes the mean square compensated distortion $d_c[n]$ as defined above.

2.2. Solution to the Minimization Problem

In this section, we determine the optimal $\boldsymbol{p}$ filter as defined in Sec. 2.1. We define the objective for optimization to be the minimum expected distortion between the compensated training and testing features, and we find $\boldsymbol{p}$ to minimize $E[d_c^2[n]]$. Using (3), obtain $E[d_c^2[n]]$ as below:

$$\begin{aligned} \overline{d_c^2} &= E[d_c^2[n]] \\ &= \sum_{0 \le i,j \le N_p-1} p[i]p[j]E[d_u[i]d_u[j]] \end{aligned} \tag{4}$$

For evaluating $\overline{d_c^2}$ in (4), the terms $E[d_u[m]d_u[n]]$ can be obtained by using (2) as below:

$$E[d_u[m]d_u[n]] = \sum_{1 \le i,j \le N_h} E[h[i]h[j]x[m-i]x[n-j]] \tag{5}$$

Further, assuming that

$$\begin{aligned} E[h[i]h[j]] &= \sigma^2\delta[i-j], \quad \sigma^2 \neq 0 \\ E[h[i]h[j]x[m]x[n]] &= E[h[i]h[j]]\ E[x[m]x[n]], \quad \forall i,j,m,n \end{aligned} \tag{6}$$

with δ being Kronecker delta, we can obtain $E[d_u[m]d_u[n]]$ in (5) as

$$E[d_u[m]d_u[n]] = N_h\sigma^2 R_x[n-m] \tag{7}$$

where R_x is the autocorrelation sequence of $\boldsymbol{x}$. Substituting (7) into (4), we obtain

$$\overline{d_c^2} = N_h\sigma^2 \sum_{0 \le i,j \le N_p-1} p[i]p[j]R_x[i-j] \tag{8}$$

We can differentiate (8) with respect to $\boldsymbol{p}$ to find the optimal $\boldsymbol{p}$ but this will result in the optimal $\boldsymbol{p}$ being $\mathbf{0}$: if all the elements in $\boldsymbol{p}$ are equal to 0, all features in $\boldsymbol{x}$ and $\boldsymbol{y_u}$ will be mapped to 0, and the mean square distortion $\overline{d_c^2}$ will always be zero as well. While this is clearly the optimal solution in the mathematical sense, it is not a useful solution. In order to avoid the degenerate solution $\boldsymbol{p} = \mathbf{0}$ we further constrain $\boldsymbol{p}$:

$$\sum_{j=0}^{N_p-1} p[j] \neq 0 \tag{9}$$

The constraint in (9) means that the $\boldsymbol{p}$ filter must have non-zero DC gain. Next, the fact that the $\boldsymbol{p}$ filter will be applied to both training and testing implies that scaling features by the same factor in both training and testing will leave the SID accuracy unchanged. This implies that we lose no generality by using the more specific constraint on $\boldsymbol{p}$

$$\sum_{j=0}^{N_p-1} p[j] = 1 \tag{10}$$

To minimize $\overline{d_c^2}$ in (8) under (10), we construct a Lagrangian optimization criterion as below:

$$\begin{aligned} \Lambda(\boldsymbol{p},\lambda) = N_h\sigma^2 &\sum_{0 \le i,j \le N_p-1} p[i]p[j]R_x[i-j] + \\ &\lambda(\sum_{j=0}^{N_p-1} p[j] - 1) \end{aligned} \tag{11}$$

Differentiating (11) with respect to $[\boldsymbol{p}, \lambda]$ and equating the differentials to zero, we can obtain the optimal $\boldsymbol{p}$ as below:

$$\begin{bmatrix} R_x[0] & R_x[1] & \ldots & R_x[N_p-1] & 1 \\ R_x[1] & R_x[0] & \ldots & R_x[N_p-2] & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ R_x[N_p-1] & R_x[N_p-2] & \ldots & R_x[0] & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix} \times \begin{bmatrix} p[0] \\ p[1] \\ \ldots \\ p[N_p-1] \\ \lambda' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \ldots \\ 0 \\ 1 \end{bmatrix} \tag{12}$$

where $\lambda' = \frac{\lambda}{2N_h\sigma^2}$. Note that the unknown in $N_h\sigma^2$ due to the reverberation filter $\boldsymbol{h}$ has been incorporated into λ'. For later reference and compactness, we write (12) equivalently as (13).

$$\Phi \begin{bmatrix} \boldsymbol{p}^T \\ \lambda' \end{bmatrix} = \mathbf{b} \tag{13}$$

We note that R_x, the autocorrelation sequence of the clean features underlying the reverberated features, is the only unknown required to find $\boldsymbol{p}$ in (13), and specifically that the optimal solution does *not* depend on the reverberation filter $\boldsymbol{h}$. We have thus designed an optimal post-filter $\boldsymbol{p}$ which is invariant to the reverberation due to $\boldsymbol{h}$ and thus far depends only on R_x. Of course, an SID system operating in a reverberant environment can observe directly the reverberated features $\boldsymbol{y_u}$, but the autocorrelation of the clean speech R_x is not directly observable. Nevertheless, we can approximate R_x as the autocorrelation sequence obtained from clean features extracted from training data:

$$R_x[m] \approx R_T[m], \quad m = 0, \cdots N_p - 2 \tag{14}$$

where $R_T[m]$ is the autocorrelation sequence obtained from clean features in training data. Combining (14) and (11), we can solve for $\boldsymbol{p}$, so the $\boldsymbol{p}$ is invariant to the underlying clean features in $\boldsymbol{x}$. The Φ matrix in (13) is Hermitian but not Toeplitz, its invertibility guarantees a solution that is both optimal and unique for $\boldsymbol{p}$. Φ was found to be invertible in SID evaluations so combining (1), (6) (9), and (14), we claim that we have developed an optimal and unique solution for $\boldsymbol{p}$ which is not only invariant to the reverberation due to $\boldsymbol{h}$ but also invariant to the underlying clean features in $\boldsymbol{x}$. This invariance greatly simplifies our SID system. We can design $\boldsymbol{p}$ using training features alone and use $\boldsymbol{p}$ to generate compensated training features $\boldsymbol{x_c}$, as in Fig. 1. The processed features $\boldsymbol{x_c}$ are used for training speaker models. During testing we apply the same filter $\boldsymbol{p}$ to the observed testing features in $\boldsymbol{y_u}$ and generate compensated testing features $\boldsymbol{y_c}$, again as in Fig. 1. Because the same $\boldsymbol{p}$ is applied across all reverberation conditions, no modification of the filter design needs to be done for any particular reverberant environment.

3. EXPERIMENTAL VERIFICATION

3.1. Experimental Procedures

We applied our reverberation compensation approach based on post-filtering to a subset of YOHO database[6]. For the YOHO dabase, a total of 40 speakers labeled from $101 - 140$ in the *enroll* part of the YOHO database were selected as subjects. SID training involved 16 utterances from $Session1$, and testing was performed with 4 utterances from $Session2$. Testing was done on clean and reverberant conditions.

Reverberant speech was obtained by convolving clean speech with simulated room impulse responses produced by the *RIR* simulator for room acoustics [7], which is based on the image method. We used a simulated room with dimensions $5 \times 4 \times 3m$, a single microphone located at the center of the room, and a distance of $1m$ between the source and the microphone. Simulated room Reverberation times (RTs) ranged from 0 to 2 s, as noted in Fig. 3. We use the standard definition for RT, the time required for the acoustic signal power to decay by 60 dB from the instant a sound source is turned off. We depict typical simulated RIRs for RTs of 0.5 and 2 seconds in Fig. 2. The larger RIRs create distortions that severely degrade SID accuracy. We applied our post-filtering algorithm to conventional 13-dimensional Mel frequency cepstral coefficients (MFCC features), as developed using the CMU SPHINX system [8]. Processing steps included silence detection, removal of the zeroth cepstral ($C0$) coefficient, and Cepstral Mean Normalization (CMN) [8]. For SID training, Gaussian mixture models (GMMs) were trained on each speaker's training data. During testing, a log-likelihood score based test was used to identify the speaker[9].

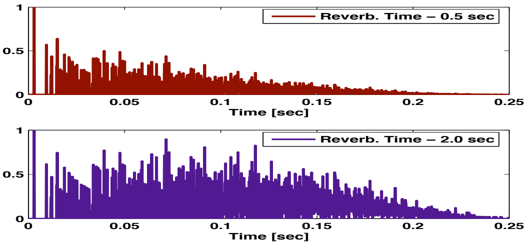

Fig. 2. Typical Room Impulse Response

3.2. Experimental Results

Using the procedures above, SID accuracy results are summarized in Fig. 3. In all cases 32 Gaussian mixtures were used, which provided our best performance for this task and database, averaged across the reverberation conditions used. The SID accuracy curve labeled "GMM" corresponds to the baseline *uncompensated* case. Compensated training features were created by applying the $\boldsymbol{p}$ filter to training features and developing GMMs for the compensated features. In presenting the compensated results in Fig. 3 we employ the notation "GMM-P-n", which refers to the use of post-processing filter $\boldsymbol{p}$ with the parameter n denoting the duration of the FIR impulse response.

Comparing the results for the "GMM" and "GMM-P-n" cases in Fig. 3, we note that our post-filtering compensation approach provides substantial improvement in SID accuracy, with greatest improvements observed for the larger RTs. Best performance was obtained for the relatively small number of five taps, in which case the relative SID average error rate decreased by 38% compared to the uncompensated case, including a relative improvement of 50% for reverberation time of 2 seconds. It is worthwhile to note that in mismatched cases (*i.e.* training on clean speech and testing on reverberated speech), the standard deviation of the SID accuracy across reverberation conditions with our compensation algorithm was only 0.65%, compared to a standard deviation of 5.2% for uncompensated features, indicating that the compensation algorithm is very robust to reverberation.

All of the results above were obtained using the unrealistic approximation that RT is independent of frequency. We performed another set of experiments using more realistic RTs that decreased with increasing frequency using realistic frequency-dependent absorption coefficients for several common building materials. These results, summarized in the upper rows of Table 1, also show a significant improvement in SID accuracy using the compensation approach described in this paper. We also applied our algorithm on SID evaluation using several real room impulse responses [10]. The corresponding results, summarized in the lower rows of Table 1, also demonstrate significant improvements using our algorithm. Additional results applying our algorithm to a subset of the Vermobil database will appear in future studies.

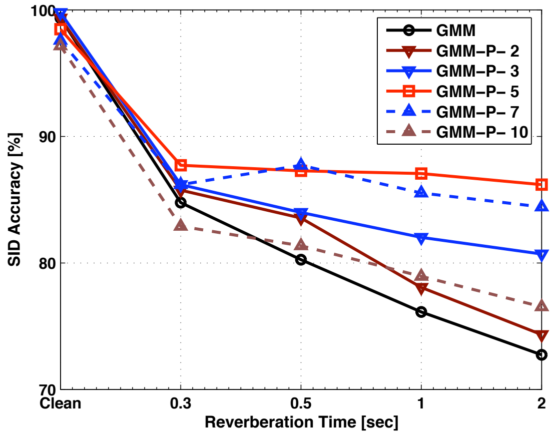

Fig. 3. SID Accuracy in Reverberation.

Table 1. SID Accuracy for Room Surfaces and real RIR

Real RIR	Room Surfaces	Uncompensated [%]	Compensated [%]
	Acoustic Tile	91.3	95.7
	Wood	71.9	86.3
	Glass	72.5	85.0
	Brick	65.6	81.9
imp-rev.30		70.0	82.5
imp-130.1		83.1	86.9
imp-160.16		97.5	97.5

4. DISCUSSION

In this section we discuss some of the assumptions made in this paper. At first we assumed a representation for reverberated features in (1). As features are generated by windowing on overlapping segments of speech signal, an exact relationship between reverberated and clean features is hard to find. We therefore, needed approximations and borrowed (1) from representation of reverberation as a LTI filer in time domain. $h_0 = 1$ was chosen to keep the problem analytically attractable.

The assumption of (6) essentially means that the frequency response of the $\boldsymbol{h}$ filter is flat. This would occur if the RTs were constant over all frequencies. This assumption was made in simulating the speech data that was used for the results in Fig. 3, but it is not empirically valid, as noted above. Nevertheless, the algorithm was also successful for the environmental conditions summarized in Table 1, which included a small number of more realistic simulated and actual room impulse responses with reverberation time that decreased as frequency increased. These data indirectly validate the proposition that the assumption of (14), although physically invalid, still can produce useful compensation results in practice.

We can obtain an estimate of number of taps in the optimal $\boldsymbol{p}$ as the knee in the curve describing the dependence of $\overline{d_c^2}$ onN_p. Although $\overline{d_c^2}$ decreases with a larger number of taps, the dissimilarity among compensated features for different speakers also decreases. As N_p becomes very large, the $\boldsymbol{p}$ filter converges to be a uniform moving average filter that smoothes out all the data and reduces every feature to its mean value, which is zero for the MFCC features under consideration. This reduces the SID decisions to a random guess. For these reasons, we expect a local maximum in performance as a function of N_p. Next, we note that the optimization construction in section 2.2 guarantees that for optimal $\boldsymbol{p}$, the mean squared distortion for compensated case $E[d_c^2[n]]$, is never greater than that for uncompensated case $E[d_u^2[n]]$. Further, the optimal $\boldsymbol{p}$ is a linear phase filter.

The post-filtering algorithm was also applied directly to the speech signal in the time domain but in this case *uncompensated* case outperformed *compensated* case. This indicates that the modeling and assumptions in Sec. 2 is not easily generalizable to the time domain.

Our approach for dereverberation is somewhat similar to Wiener Filtering at a conceptual level. The major digression from the Wiener filter is that the $\boldsymbol{p}$ is applied to both training and testing speech, which leads to different requirements and solutions to the problem. While our approach is invariant to the detailed nature of the reverberation this is not the case for Wiener Filtering. We will consider generalizations of our approach in later studies. We will also apply the algorithm for speaker verification tasks.

5. CONCLUSIONS

We presented an algorithm for reverberation compensation for SID applications. The compensation procedure consisted of a relatively simple FIR filter that is applied to sequences of cepstral coefficients, with the coefficients of the filter optimized to minimize the mean square difference between the compensated coefficients for speech in the training and testing environments. The optimal filter obtained was unique and invariant to the environmental conditions of a particular test trial. This approach provided significant improvement in SID accuracy across different reverberation conditions encompassing simulated as well as actual RIRs and also across different speech databases.

6. REFERENCES

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437–1462, september 1997.

[2] B. Xiang, U. V. Chaudhari, J. Navratil, G. Ramaswamy, and R. A. Gopinath, "Short-time gaussianization for robust speaker verification," *IEEE ICASSP*, pp. 681–684, 2002.

[3] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, No.7, pp. 2023–2032, 2007.

[4] M. L. Seltzer, *Microphone Array Processing for Robust Speech Recognition*, Ph.D. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, July 2003.

[5] J. S. Gammal and R. A. Goubran, "Combating reverberation in speaker verification," *Instrumentation and Measurement Technology Conference*, 2005.

[6] J.P. Campbell and D.A. Reynolds, "Corpora for the evaluation of speaker recognition systems," *IEEE ICASSP*, 1999.

[7] S. G. McGovern, "A model for room acoustics," `http://2pi.us/rir.html`.

[8] CMU Sphinx Open Source Speech Recognition Engines, `http://cmusphinx.sourceforge.net/html/cmusphinx.php`.

[9] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, August.

[10] RWCP Sound Scene Database in Real Acoustical Environments, `http://tosa.mri.co.jp/sounddb/indexe.htm`.