MULTISENSOR DYNAMIC WAVEFORM FUSION

Alan McCree, Kevin Brady, and Thomas F. Quatieri

MIT Lincoln Laboratory
Lexington, MA 02420
E-mail: [mccree, kbrady, tfq]@ll.mit.edu

ABSTRACT

Speech communication is significantly more difficult in severe acoustic background noise environments, especially when low-rate speech coders are used. Non-acoustic sensors, such as radar sensors, vibrometers, and bone-conduction microphones, offer significant potential in these situations. We extend previous work on fixed waveform fusion from multiple sensors to an optimal dynamic waveform fusion algorithm that minimizes both additive noise and signal distortion in the estimated speech signal. We show that a minimum mean squared error (MMSE) waveform matching criterion results in a generalized multichannel Wiener filter, and that this filter will simultaneously perform waveform fusion, noise suppression, and crosschannel noise cancellation. Formal intelligibility and quality testing demonstrate significant improvement from this approach.

Index Terms— Non-acoustic sensor, waveform fusion

1. INTRODUCTION

Since speech coding in severe acoustic noise is very difficult, recent work has explored the use of non-acoustic sensors to supplement the acoustic microphone information [1, 2, 3]. In particular, waveform fusion of non-acoustic sensors, combined with additional highband speech encoding, produced significant intelligibility improvements for the 2.4 kb/s NATO MELPe standard [3].

In this work, we have extended this earlier fixed waveform fusion approach to a dynamic waveform fusion algorithm that combines sensor fusion, noise suppression, and crosschannel noise cancellation into a single time-varying filter. This algorithm is an extension of the multichannel Wiener filtering approach to incorporate both additive noise and signal distortion.

2. DYNAMIC WAVEFORM FUSION

We have developed a dynamic waveform fusion algorithm to combine acoustic and non-acoustic sensors. This approach uses a MMSE criterion for optimization, incorporating both additive noise and signal distortion.

2.1. Preliminaries: Review of Wiener Filter

To introduce our notation, as well as to provide a starting point for further analysis, we begin with a review of the well-known Wiener Filter.

This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2.1.1. Single Channel

If we observe the noisy time signal $y(t) = s(t) + n(t)$, where $s(t)$ and $n(t)$ are the signal and additive noise, respectively, then in the frequency domain we have $Y(f) = S(f) + N(f)$. We wish to find the optimal linear filter $G(f)$ to estimate the signal from the noisy observation, i.e. $\hat{S}(f) = G(f)Y(f)$. Dropping the frequency index f for convenience and using the inner product notation $\langle A, B\rangle$ to represent the expected value of A^*B and $\|A\|^2 = \langle A, A\rangle$, we can write the mean-squared error as

$$E = \|S - \hat{S}\|^2 = \|S - GY\|^2.$$

By taking the derivative with respect to G, we get the MMSE solution:

$$\begin{aligned} G_{opt} &= \frac{\langle S, Y\rangle}{\|Y\|^2} \\ &= \frac{\|S\|^2 + \langle S, N\rangle}{\|S\|^2 + \|N\|^2 + \langle S, N\rangle} \end{aligned}$$

Assuming the signal and noise to be independent gives the familiar result:

$$G_{opt} = \frac{\|S\|^2}{\|S\|^2 + \|N\|^2} = \frac{P_s}{P_s + P_n}, \tag{1}$$

where P_s and P_n represent the signal and noise power at frequency f. In most applications, the noise is assumed to be stationary, so that P_n can be estimated during silent periods. P_s can also be a stationary signal power estimate, or it can be dynamically estimated for each speech frame in which case the resulting Wiener filter will be time-varying.

2.1.2. Multichannel Wiener Filter

For multiple acoustic microphones, the Wiener filter can be generalized. Given two channels, Y_1 and Y_2, such that $Y_1 = S + N_1$ and $Y_2 = S + N_2$, the MMSE solution for waveform combination $\hat{S} = G_1Y_1 + G_2Y_2$ is achieved with coefficients defined by:

$$\begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} \|Y_1\|^2 & \langle Y_2, Y_1\rangle \\ \langle Y_1, Y_2\rangle & \|Y_2\|^2 \end{bmatrix}^{-1} \begin{bmatrix} \langle S, Y_1\rangle \\ \langle S, Y_2\rangle \end{bmatrix}$$

In vector notation, we have

$$\mathbf{G}(f) = \mathbf{R}_{yy}^{-1}(f)\mathbf{R}_{sy}(f) \tag{2}$$

with

$$\hat{S}(f) = \mathbf{G}^T(f)\mathbf{Y}(f). \tag{3}$$

In the more general case, each actual observation Y_i^o has passed through a transfer function H_i:

$$Y_i^o = H_iY_i = H_i(S + N_i).$$

Each channel should then be equalized prior to Wiener filtering, using

$$Y_i = H_i^{-1} Y_i^o.$$

Each transfer function represents the signal path to the sensor. In our application, all sensors are mounted to the talker, so the transfer functions should not vary over time. Therefore, the equalization functions H_i^{-1} can be measured using a microphone calibration process in the absence of background noise. If we use a high-quality reference microphone to record the clean signal S (or simply use the resident acoustic microphone as the true signal), then for each sensor:

$$H_i^{-1} = \frac{\langle S, Y_i^o \rangle}{\|Y_i^o\|^2}. \tag{4}$$

In the absence of noise, this is an exact relationship. If a quiet calibration process is infeasible, or if the transfer functions may drift with time, the equalization functions can be estimated from time intervals where the speech is much louder than the background noise. In all subsequent analysis, we will assume that this equalization has been done prior to waveform fusion.

2.2. Signal Distortion

For nonacoustic sensors, additive noise is not the primary signal corruption. These transducers do not have full response to the speech signal, especially at higher frequencies. As a result, the nonacoustic sensors of interest to us generate significant signal distortion after equalization, even in the absence of background noise. If we model this distortion as a time-varying transfer function noise, then we get both multiplicative and additive noise terms in the frequency domain:

$$Y_i(f) = [1 + H_{di}(f)]S_i(f) + N_i(f),$$

where H_{di} represents the unknown component of the overall transfer function. Note that if this tranfer function noise is independent of the signal, we can also view this as

$$Y_i(f) = S_i(f) + D_i(f) + N_i(f)$$

where $D_i(f)$ is an additive distortion with variance proportional to the signal variance.

Like the transfer functions, this distortion power can be measured in a quiet signal calibration process, in this case based on the coherence between the reference microphone and the sensor in question. The signal coherence for channel i is given by:

$$\rho_i = \frac{\langle S, Y_i \rangle}{\sqrt{\|S\|^2 \|Y_i\|^2}}. \tag{5}$$

For the calibration process there is no background noise, so $Y_i = S + D_i$, and the coherence ρ_i is only reduced by signal distortion:

$$\rho_i = \frac{\|S\|^2 + \langle S, D_i \rangle}{\sqrt{\|S\|^2(\|S\|^2 + \|D_i\|^2 + \langle S, D_i \rangle)}}.$$

Assuming the signal and distortion to be uncorrelated (since any component of distortion that is linearly related to the signal would be incorporated in the equalization filter) leads to:

$$P_{di} = P_s \left(\frac{1}{\rho_i^2} - 1 \right) \tag{6}$$

where P_{di} is the distortion power at this frequency. Note that for high-quality sensors, such as close-talking acoustic microphones, the calibration coherency should be equal to 1 so that the distortion power will be zero.

2.3. Static Waveform Fusion

Our initial work with MMSE waveform fusion was to find the optimal static fusion coefficients for a given recording. To use the Wiener filter in Eq. 2, we must estimate $\mathbf{R}_{yy}(f)$ and $\mathbf{R}_{sy}(f)$. We make the following assumptions:

1. The signal, noise, and distortion are stationary.
2. The noise is independent of the signal.
3. The noises in different channels are independent.
4. The distortion is independent of the signal.
5. The distortions in different channels are independent.
6. The noise and distortion are independent.

Then all cross-terms vanish, leading to the following expressions (for the two channel case):

$$\mathbf{R}_{sy} = \begin{bmatrix} P_s \\ P_s \end{bmatrix} \tag{7}$$

and

$$\mathbf{R}_{yy} = \begin{bmatrix} P_s + P_{n1} + P_{d1} & P_s \\ P_s & P_s + P_{n2} + P_{d2} \end{bmatrix} \tag{8}$$

We have implemented a waveform fusion algorithm using this approach. We estimate the necessary speech, noise, and distortion parameters in each channel using a sequence of Discrete Fourier Transforms (DFTs). First, the channels are equalized to the reference microphone using Eq. 4 from the quiet calibration recordings. Then, the speech power for frequency f is estimated as the mean power in the primary microphone at that frequency, the noise power is estimated as the 20th percentile point of the total signal power at frequency f in channel i, and the distortion power for f and i is estimated using Eq. 6.

For sensor configurations similar to those in [3], this results in comparable performance gain. The advantage of this approach is that it automatically adjusts the tradeoff between additive noise and signal distortion at each frequency based on the relative performance of each sensor with the current noise environment and talker.

2.4. Dynamic Fusion

The assumption that the speech signal is stationary may be unnecessarily restrictive. A more general case is to replace the first assumption from the previous section with the following:

- The noise is stationary and the signal-to-distortion ratio is constant.

The resulting waveform fusion will be *dynamic*, since the coefficients will change from frame to frame. This does not require any change to the filter estimation equations, simply that the speech power is estimated based on each frame's DFT. As in Wiener filter-based noise suppression algorithms, this speech power can be estimated by

$$P_s = P_y - P_n$$

from the primary acoustic microphone.

Unfortunately, initial experiments with this straightforward approach show that it results in an annoying tonal background noises, similar to the well-known "musical noise" generated by noise suppression algorithms. Experimental analysis shows that the time-varying Wiener filter is indeed acting like a noise suppression algorithm, since the sum of the fusion coefficients is often much less than one.

2.4.1. Disabling Noise Suppression

One approach to eliminating the musical noise is to constrain the fusion algorithm not to attenuate the signal. This can be done using a Lagrangian technique, with a constraint on the sum of the coefficents:

$$E = \|S - \hat{S}\|^2 + 2\lambda\|S\|^2 \left|1 - \sum_i G_i^c\right|$$

This leads to the partial solution for the optimal constrained coefficients:

$$\mathbf{G}^c(f) = (1 - \lambda(f))\mathbf{R}_{yy}^{-1}(f)\mathbf{R}_{sy}(f)$$

Imposing the constraint:

$$1 = \mathbf{1}^T\mathbf{G}^c(f) = (1 - \lambda(f))\mathbf{1}^T\mathbf{R}_{yy}^{-1}(f)\mathbf{R}_{sy}(f)$$

Defining the sum of the unconstrained Wiener filter coefficients as

$$G_{sum}(f) = \mathbf{1}^T\mathbf{R}_{yy}^{-1}(f)\mathbf{R}_{sy}(f),$$

we can solve for λ

$$(1 - \lambda(f))G_{sum}(f) = 1$$

$$\lambda(f) = \frac{G_{sum}(f) - 1}{G_{sum}(f)}$$

so that

$$\mathbf{G}^c(f) = \frac{1}{G_{sum}(f)}\mathbf{G}(f) \tag{9}$$

where $\mathbf{G}(f)$ is given by Eq. 2. This result is intuitively satisfying: the optimal gain-constrained filter is the optimal unconstrained filter divided by the sum of the unconstrained coefficents.

In informal evaluations, this constrained waveform fusion approach does remove the musical noise distortion. Unfortunately, this also reduces the amount of noise reduction. While it is possible to run a separate noise suppression algorithm on the waveform fusion output, it would be more satisfying to improve the unconstrained fusion algorithm to retain the noise suppression functionality without introducing tonal artifacts.

2.4.2. Robust Signal Power Estimation

Our efforts to increase the robustness of the algorithm have focussed on the estimation of the signal power P_s. In noise suppression algorithms, it is well-known that the musical noise phenomenon is related to measurement fluctuations in the instantaneous estimation of the signal power, and that this can be addressed by smoothing the estimates either across frequency [4] or time [5]. We generalize this concept to our multichannel waveform fusion algorithm.

We have developed a three-stage approach to robust signal power estimation. The first estimate uses only the resident acoustic microphone: $P_s^{(1)} = P_y - P_n$. With this initial estimate of P_s, we apply gain-constrained waveform fusion to produce a second, multichannel estimate of the signal power. Since

$$\hat{S} = \sum_i G_i^c Y_i = S\sum_i G_i^c + \sum_i G_i^c N_i = S + \sum_i G_i^c N_i,$$

then

$$S = \sum_i G_i^c (Y_i - N_i),$$

Therefore, our second estimate of P_s is given by

$$P_s^{(2)}(f) = (\mathbf{G}^c)^T(f)(\mathbf{R}_{yy} - \mathbf{R}_{nn})\mathbf{G}^c(f).$$

Finally, we incorporate time smoothing using the decision-directed *a priori* estimator of Ephraim and Malah, where the current frame signal power is smoothed with the previous filter output [6, 5]:

$$P_s^{(3)} = (1 - \alpha)P_s^{(2)}(f) + \alpha|\hat{S}_{prev}|^2$$

where $\hat{S}_{prev}(f)$ is the waveform fusion output from the previous frame and α is a smoothing parameter (typically 0.98). Using $P_s^{(3)}$ in Eq. 7, 8, and 2 results in a waveform fusion with strong noise suppression but minimal noise artifacts and signal distortion.

2.4.3. Cross-channel Noise Cancellation

Typically, with spatially-diverse noise sources and sensors, the noise is essentially independent across channels. However, in special cases, such as when two acoustic sensors are deliberately positioned close to each other, there may be coherency between the noise components in these sensors. In these scenarious, our third assumption in Sec. 2.3 is not reasonable. Estimating a full noise covariance matrix, rather than assuming non-diagonal terms to be zero, results in the following:

$$\mathbf{R_{yy}} = \begin{bmatrix} P_s + P_{n1} + P_{d1} & P_s + \langle N_2, N_1\rangle \\ P_s + \langle N_1, N_2\rangle & P_s + P_{n2} + P_{d2} \end{bmatrix} \tag{10}$$

As in the simpler prior noise estimation algorithm, these noise cross-terms can be estimated during silent periods.

With this approach, the waveform fusion algorithm implements an additional *noise cancellation* functionality by exploiting noise coherency across channels to remove noise components. Note that this also results in complex fusion coefficients $\mathbf{G}(f)$, which are purely real for all prior examples.

3. PERFORMANCE EVALUATION

We have tested this dynamic waveform fusion algorithm with a set of six sensor signals. These sensors are a dual channel close-talking noise cancelling microphone from Aliph Corporation, two channels of a second-generation microwave radar sensor mounted on the throat (also from Aliph), a piezo-electric vibrometer also on the throat (P-mic), and a bone conduction microphone located on the top of the skull (bone-mic). For comparison, a typical resident microphone (Gentex M175A) was also available.

3.1. Testing With Lombard Speech in Noise

For initial testing of the dynamic waveform fusion algorithm, we digitally mixed clean multichannel recordings with separate multichannel noise recordings. The clean speech signal was produced with an induced Lombard effect to better simulate speech in noise. This testing method allows us to compare the clean speech input, noisy mixed signals, and fusion output. Each signal was coded with the 2.4 kb/s NATO MELPe coder [7], and the ITU PESQ objective measure was used for evaluation. The test involved a total of 12 files, each 20 seconds long, from six speakers in two noise fields: Blackhawk helicopter and Bradley fighting vehicle. These results are shown in Table 1. The baseline system has a predicted MOS of 2.63. Microphone equalization and noise suppression, achieved by running the dynamic waveform fusion software with only the resident microphone as input, results in a small improvement in performance. Static waveform fusion using all six sensors provides more improvement, but the dynamic fusion performs best. This system approaches but does not fully achieve the performance of MELPe with the clean reference microphone input.

Method	Score
Resident mic.	2.63
Resident equalized	2.68
Static fusion	2.77
Dynamic fusion	3.01
Clean ref. mic.	3.25

Table 1. PESQ scores of waveform fusion algorithms with 2.4 kb/s MELPe coding.

Method	DRT
Resident equalized	86.5
Dynamic fusion	91.9

Table 2. Intelligibility comparison of dynamic waveform fusion vs. equalized resident microphone signal with 6 kHz bandwidth unquantized MELP coding in Blackhawk environment.

3.2. Formal Testing

We also compared the six-sensor dynamic waveform fusion against the resident microphone in a Diagnostic Rhyme Test (DRT). We used six talkers in the Blackhawk helicopter environment, with the fusion outputs encoded by an unquantized MELP system using 6 kHz bandwidth. These results, shown in Table 2, also show significant performance improvement from the fusion system as compared to the equalized resident microphone.

Finally, we performed a quality evaluation using an A/B test between the equalized resident microphone and the dynamic fusion algorithm. Again, both signals were encoded by a 6 kHz unquantized MELP system, and the Blackhawk helicopter environment was used. Preference scores, averaged over 8 speakers, are shown in Table 3. This improvement is statistically significant with a 95% confidence interval.

4. CONCLUSION

By incorporating a multiplicative noise model into a multichannel Wiener filtering approach, we have shown that non-acoustic signals can be optimally exploiting using a MMSE criterion. This approach results in automated static waveform fusion that is appropriate for a particular talker and environment. For dynamic fusion, we have developed an additional algorithm for estimating the instantaneous SNR. Formal testing results show that the resulting dynamic waveform fusion algorithm provides significant intelligibility and quality improvement for low-rate coding in difficult acoustic environments.

Method	Preference
Resident equalized	37%
Dynamic fusion	63%

Table 3. Quality comparison of dynamic waveform fusion vs. equalized resident microphone signal with 6 kHz bandwidth unquantized MELP coding in Blackhawk environment.

5. ACKNOWLEDGEMENTS

We thank John Tardelli and Paul Gatewood from the ARCON Corporation for multisensor corpora as well as DRT and A/B test results, and Alan Curtis of BBN for additional corpus assistance. Thanks also to Sam Earp for interesting discussions on multiplicative noise modeling.

6. REFERENCES

[1] L. C. Ng, G. C. Burnett, J. F. Holzrichter, and T. J. Gable, "Denoising of Human Speech Using Combined Acoustic and EM Sensor Signal Processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 229–232.

[2] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, "Multi-sensory Microphones for Robust Speech Detection, Enhancement and Recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004, pp. 781–784.

[3] T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting Nonacoustic Sensors for Speech Encoding," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 533–544, Mar. 2006.

[4] L. Arslan, A. McCree, and V. Viswanathan, "New Methods for Adaptive Noise Suppression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Detroit, 1995, pp. 812–815.

[5] O. Cappe, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[6] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[7] A. McCree and T. P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.