A Performance Monitoring Approach to Fusing Enhanced Spectrogram Channels in Robust Speech Recognition

Shirin Badiezadegan and Richard Rose

Department of Electrical and Computer Engineering, McGill University, Montreal, Canada

shirin.badiezadegan@mail.mcgill.ca, rose@ece.mcgill.ca

Abstract

An implementation of a performance monitoring approach to feature channel integration in robust automatic speech recognition is presented. Motivated by psychophysical evidence in human speech perception, the approach combines multiple feature channels using a closed loop criterion relating to the overall performance of the system. The multiple feature channels correspond to an ensemble of reconstructed spectrograms generated by applying multiresolution discrete wavelet transform analysis-synthesis filter-banks to corrupted speech spectrograms. The spectrograms associated with these feature channels differ in the degree to which information has been suppressed in multiple scales and frequency bands. The performance of this approach is evaluated in both the Aurora 2 and the Aurora 3 speech in noise tasks.

Index Terms: spectrographic mask, wavelet-based de-noising, spectrogram reconstruction

1. Introduction

There are many examples of very low signal-to-noise ratio environments and communications channels that render automatic speech recognition (ASR) systems nearly inoperable when applied to speech utterances taken from these domains. The family of imputation based missing feature (MF) techniques attempts to reduce the impact of these environments by characterizing the interaction between speech and acoustic background spectral components. This information is then used to reconstruct the underlying uncorrupted spectral information in the speech utterance. The information concerning this interaction is acquired from separately estimated spectro-temporal masks which provide either discrete thresholds or continuous probabilities to indicate the presence of speech in time-frequency spectral bins.

The interest in this work is to generalize these techniques to generate multiple parallel reconstructed spectrograms where each spectrogram corresponds to one of an ensemble of thresholding schemes. The parallel channels associated with these spectrograms are combined to generate ASR features using a closed-loop “performance monitoring” approach to combining multiple channels. This general class of approaches, as originally proposed in [1], is motivated by psychophysical evidence in human perception [2]. The evidence supporting this performance monitoring approach in human perception suggests that listeners may be able to suppress information from unreliable perceptual channels and select reliable channels based on their assessment of whether a message has been reliably received [2].

The reconstructed spectrograms which form the parallel channels in this performance monitoring approach are obtained using a multi-resolution discrete wavelet transformation (DWT) approach to spectrogram reconstruction [3]. This is a missing feature based approach that was motivated by theory arising from wavelet-based de-noising [4]. Noise corrupted spectrograms from a speech utterance are presented to a pyramidal wavelet-based analysis-synthesis filter-bank. Thresholds are derived from a spectrographic mask to suppress noise corrupted wavelet coefficients at multiple scales and frequency bands in the analysis filter-bank. In this sense, the re-synthesized spectrogram components produced by the synthesis filter-bank have been “de-noised” according to a given thresholding strategy. A brief description of the approach is provided in Section 2.

The premise of the work described in this paper is that an ensemble of thresholding strategies can be posed for suppressing wavelet coefficients, each providing different detection characteristics at different filter-bank scales and frequency bands. This ensemble of thresholding strategies results in an ensemble of reconstructed spectrograms, each differing in the degree to which information has been suppressed in scale and frequency. It is this ensemble of spectrograms that form the parallel processing channels in the performance monitoring approach described in Figure 1 and presented in Section 3.

Figure 1: Block diagram of the performance monitoring approach to channel fusion.

The overall approach presented here consists of three major steps and is summarized in the block diagram shown in Figure 1. The first step is the extraction of parallel feature channels, \( S = \{S_1, \ldots, S_M\} \), from the corrupted speech. These channels correspond to the reconstructed spectrograms associated with the threshold strategies \( \Omega = \{\Omega_1, \ldots, \Omega_M\} \) described above. The second step in the performance monitoring approach is to combine these channels according to a channel fusion function, \( f_\alpha(S) \), where \( \alpha \) represents the parameters associated with the channel fusion. For each possible setting of \( \alpha \), a symbol sequence \( R^C_\alpha = r^C_1, \ldots, r^C_N \) is generated for the
N frame corrupted speech utterance. Following the procedure suggested in [1], each \( r_{n} \) is a vector of phone posteriors generated by a neural network trained to predict sub-word symbols from input channels.

Finally, the last step is to derive sample statistics \( C_{\alpha}^{G} \) from corrupted speech observation sequence. The goal is to determine the degree to which these sample statistics deviate from sample statistics, \( C_{\alpha}^{G} \), derived from similar sequences previously observed under uncorrupted conditions. The optimum fusion strategy is determined by the fusion parameters, \( \alpha = \text{argmax}_{\alpha} \{ \Phi(f_{n}, C_{\alpha}^{G}) \} \).

2. Wavelet-based Spectrogram Reconstruction

A multi-resolution DWT approach to spectrogram reconstruction for robust ASR was originally presented in [3]. Section 2.1 briefly introduces this approach as a means for masking wavelet coefficients by exploiting speech presence probability (SPP) estimates obtained from spectrographic masks. Section 2.2 describes how the wavelet-domain mask is used for selective wavelet reconstruction.

2.1. Generating wavelet-domain masks

The spectrogram reconstruction approach in [3] is motivated by the theoretical arguments for wavelet based signal de-noising originally presented by Donoho in [4]. These methods have been shown to achieve near optimal estimates of the original signal from noisy observations when wavelet coefficients are thresholded using an “oracle” thresholding scheme [4]. This thresholding scheme identifies and preserves wavelet coefficients generated from additive corrupting noise. In a missing data framework, this process can be rephrased as identifying the “reliable” and “unreliable” wavelet coefficients of the noise corrupted speech frame and performing the wavelet-based masking, accordingly.

In DWT-based spectrogram reconstruction, the noisy speech spectrogram is presented to the DWT filter-bank and thresholds are estimated for wavelet coefficients at all filter-bank scales [3]. At each analysis frame, a D-dimensional vector of log energy coefficients, \( y = [y_{1}, \ldots, y_{D}] \), is extracted from noise corrupted speech. A speech presence probability (SPP) vector \( \theta = [\theta_{1}, \ldots, \theta_{2}] \) is estimated from a spectrographic mask. Each \( \theta_{j} \) represents the probability that underlying speech spectral energy has not been masked by noise. The next step is to obtain the mask values at each filter-bank scale from \( \theta \) by propagating \( \theta \) through the DWT filter bank [3]. The mask propagation begins at the first wavelet scale where the mask vector, \( \Theta_{1} = [\Theta_{1,1}, \ldots, \Theta_{1,K_{1}}] \), for the first scale is thresholded and applied to the wavelet coefficients \([Y_{1,1}, \ldots, Y_{1,K_{1}}]\) in which \( K_{1} \) is the number of wavelet coefficients at the first scale. A similar approach is taken to create a wavelet-domain mask vector at the first scale for approximation coefficients of \( y \), \([A_{1,1}, \ldots, A_{1,K_{1}}]\), resulting in mask components \( \Delta_{1} = [\Delta_{1,1}, \ldots, \Delta_{1,K_{1}}] \) for the approximation coefficients. This process can be repeated so that the mask vectors at the \( j \)th scale, \( \Theta_{j} \), and \( \Delta_{j} \), are propagated to the mask vectors at the \( j + 1 \)st scale, \( \Theta_{j+1} \) and \( \Delta_{j+1} \), up to the \( J \)th scale in the DWT filter-bank.

2.2. Selective wavelet reconstruction

The last step in the wavelet de-noising process is to generate a binary mask to be applied to wavelet coefficients, \( Y_{j,k} = [Y_{j,1,k}, \ldots, Y_{j,K_{j},k}] \), at each of \( j = 1, \ldots, J \) scales. The binary mask, \( \hat{\Theta}_{j,k} \), for wavelet coefficient, \( Y_{j,k} \), is obtained from the continuous mask value, \( \Theta_{j,k} \) by applying a threshold \( \Lambda_{j} \):

\[
\hat{\Theta}_{j,k} = \begin{cases} 
1 & \text{if } \Theta_{j,k} \geq \Lambda_{j} \\
0 & \text{else}
\end{cases}
\]

(1)

A similar approach is performed to map the continuous-valued elements of \( \Delta_{j} \) to binary-valued components of \( \hat{\Delta}_{j} \) using the threshold values \( \Gamma_{1}, \ldots, \Gamma_{J} \). Hence, a set of \( 2J \) threshold values \( \Omega = \{ \Lambda_{1}, \ldots, \Lambda_{J}, \Gamma_{1}, \ldots, \Gamma_{J} \} \) are needed to specify the de-noising strategy. Having determined the binary wavelet domain masks at each of \( J \) scales, they are used to mask the wavelet coefficients:

\[
y_{j,k}^{\text{hard}} = \begin{cases} 
Y_{j,k} & \text{if } \hat{\Theta}_{j,k} = 1 \\
0 & \text{else}
\end{cases}
\]

(2)

Since the approximation coefficients are the outputs of the low-pass filters in the DWT pyramidal filter-bank, corrupting noise has the effect of introducing slowly varying components into these coefficients. To deal with this type of corruption, “unreliable” approximation coefficients are smoothed with the adjacent “reliable” coefficients. Having masked the wavelet coefficients and the performed smoothing on the approximation coefficients, the inverse discrete wavelet transform is applied to reconstruct the log mel-spectral features, \( \tilde{y} \).

3. Channel Fusion

This section describes the “channel fusion” portion of the performance monitoring system depicted in Figure 1. The optimum combination of feature channels is obtained by generating observation sequences from the combined feature channels and evaluating the similarity of these sequences with respect to sequences generated under uncorrupted conditions. The feature channels, \( S = \{S_{1}, \ldots, S_{M}\} \), correspond to \( M \) possible detection threshold settings \( \Omega = \{\Omega_{1}, \ldots, \Omega_{M}\} \) for selective wavelet reconstruction as described in Section 2.2.

The observation vectors in Figure 1 correspond to posterior probabilities of observing sub-word units \( p_{n,l} \), \( l = 1, \ldots, L \) given the fusion of the input channels, \( f_{n}(S) \). Hence, component \( l \) of the observation vector \( r_{n} \) for frame \( n \) in an \( N \) frame utterance is given as \( r_{n}^{\text{sub}}[l] = P(p_{n,l} f_{n}(S)) \). For the special case where \( \alpha \) corresponds to a binary selection of input channels, this component of the observation vector is written as \( r_{n}^{\alpha}[l] = P(p_{n,l} | S_{\alpha}[n]) \) where \( S_{\alpha}[n] \) is the vector of reconstructed spectrogram components for the \( n \)th frame of the \( \alpha \)th channel.

A single layer multilayer perceptron is trained from clean speech utterances to generate these vectors of posterior probabilities. A set of \( L = 12 \) sub-word units are defined by clustering the states of the hidden Markov models (HMM) representing the eleven words in the Aurora digit recognition task described in Section 4. The output activations of the neural network at frame \( n \) are estimates of the posterior probabilities for the \( L \) sub-word classes and the inputs are vectors of spectrogram components for that frame.

To assess the quality of feature channel \( S_{\alpha}[n] \), for an \( N \) frame utterance, the first step is to estimate the sequence of posterior
The accumulation of the probability vectors \( R_{C_m}^C = [r_1^{C_m}, \ldots, r_N^{C_m}] \). The long-term statistics of this sequence are described by the auto-correlation matrix \( C_{\alpha m}^C = \sum_{n=1}^{N} \left( \hat{r}_m^n \right)^T \).

The accumulation of the \( N \times N \) matrix, \( C_{\alpha m}^C \), can be interpreted as an unsupervised means for characterizing inter-symbol confusions between decoded sub-word symbol sequences. The magnitude of the diagonal elements of \( C_{\alpha m}^C \) is proportional to the level of confidence for the NN in individual sub-word units and the magnitude of the off-diagonal elements is proportional to the uncertainty associated with pairs of sub-word units. It should be expected then that this matrix should be more diagonally dominant when the \( R_{C_m}^C \) are obtained from uncorrupted utterances.

An autocorrelation matrix, \( C^U \), is computed from approximately 150,000 frames of the uncorrupted training speech data to serve as a reference in the closed loop system. A similarity measure is defined for measuring the degree to which \( C_{\alpha m}^C \) deviates from the uncorrupted speech autocorrelation matrix, \( C^U \).

For corrupted utterances, the feature channels generating observations where \( C_{\alpha m}^C \) is more similar to the reference matrix, \( C^U \), are potentially the channels with the least deviation from uncorrupted conditions. The similarity measure used for evaluating this deviation is given by

\[
\Phi_m = \sum_i \sum_j \left[ C_{\alpha i m}^C \right]_{i,j} \left[ C^U \right]_{i,j},
\]

which provides a point-wise 2-dimensional correlation measure. Feature channels with higher values of \( \Phi_m \) correspond to reconstructed spectrograms that are more similar to those extracted from uncorrupted training features.

The similarity measures \( \Phi_1, \ldots, \Phi_\alpha \) are fed back to the fusion block to determine the fusion parameters, \( \alpha \). For the experiments described in Section 4, the channels with the highest \( \Phi \) values are either selected as a single best channel or combined in a weighted sum of channels with weights proportional to their \( \Phi \) values. In our implementations, we select the channel with the maximum \( \Phi \) value, \( \Phi_{\text{max}} \), if there is a significant difference between \( \Phi_{\text{max}} \) and the next highest \( \Phi \). Otherwise, the optimum reconstructed spectrogram is computed from the weighted sum of the three channels with the highest \( \Phi \) values. While this is one of many possible approaches to channel fusion, ongoing research is being directed towards developing more powerful and efficient channel fusion strategies.

### 4. Experimental Study

This section describes the experimental study conducted to evaluate the performance of the performance monitoring approach presented in Sections 2 and 3 for both the Aurora 2 and the Aurora 3 noisy speech tasks. The study will compare the performance of this multi-channel closed loop approach with an implementation of a missing feature based minimum mean squared error (MMSE) approach for spectrogram reconstruction [5, 6].

#### 4.1. Task domain and implementation

All approaches were evaluated on the Aurora 2 (test set "a") and Aurora 3 (Spanish dataset) speech in noise connected digit task domain. The Spanish Aurora 3 corpus was collected in a car environment under multiple driving conditions using both close-talking and far-field microphones. Three scenarios were used for training and testing where each scenario differs in the degree of mismatch between training and testing conditions. The highest of the three mismatch conditions were used in the experiments described in Section 4.2. ASR feature analysis was performed by extracting mel log spectral features using a 25 ms Hamming window, updated every 10 ms. A 512-point FFT was applied to evaluate the spectral values, and a mel-scale filter-bank with D=23 filters was used to generate the log mel-spectral features over a 4000 Hz bandwidth. Whole word digit models were trained using the Aurora 2 and Aurora 3 clean speech training sets.

The DWT based analysis-synthesis filter-bank discussed in Section 2 is implemented using a symlet 4 wavelet basis which has previously been used in a speech de-noising application [7]. A \( J = 4 \) level filter-bank structure was found to provide sufficient resolution when evaluating the filter-bank for the single-channel spectrogram reconstruction techniques presented in [3] and is also used for the multi-channel systems in this work.

A long term goal of this work is to develop a formalism for determining a set of feature channels that are sufficient for preserving underlying speech information in the presence of interfering distortions. However, in this work, an ad-hoc strategy is used for specifying a set of \( M = 8 \) channels of reconstructed spectrograms where the \( m \)th channel is specified by the threshold values, \( \Omega_m = \{ \Lambda_1^m, \ldots, \Lambda_4^m, \Gamma_1^m, \ldots, \Gamma_4^m \} \), described in Section 2. Hence, the \( 8 \) values for each of \( 8 \) sets of \( \Omega_m \) thresholds must be determined to specify the de-noising strategies for the set of \( 8 \) channels. Each threshold value can take on either a "high" or "low" level where the actual levels are determined from development utterances by observing the approximate distributions of the wavelet coefficient masks, \( \Theta_{j,k} \), and the approximation coefficient masks, \( \Delta_{j,k} \). For a given channel that performs the DWT based analysis-synthesis at scale \( j \), \( 1 \leq j \leq 4 \), all the threshold values \( \{ \Lambda_1^m, \ldots, \Lambda_4^m, \Gamma_1^m, \ldots, \Gamma_4^m \} \) are either "high" or "low" and the remaining threshold values which correspond to higher scales, \( \{ \Lambda_5^m, \ldots, \Lambda_8^m, \Gamma_5^m, \ldots, \Gamma_8^m \} \), are set to zero. This strategy limits the number of channels.

#### 4.2. ASR performance

Table 1 displays the ASR word accuracy (WAC) for all systems evaluated on the Aurora 2 corpus for four different noise types and four SNRs ranging from 5dB to 20dB. Four separate tables are displayed, one for each noise type (subway, speech babble, car environment, and exhibition hall) where each of these tables contains ASR WACS for four different systems. For each of these tables, the first row displays the baseline ASR WAC obtained when no feature compensation is performed. The second row, labeled “MMSE”, corresponds to an implementation of a well-known missing feature based minimum mean squared error approach to spectrogram reconstruction [5, 6]. A discussion of this implementation is presented in [6]. The third row in these tables, labeled “single channel”, corresponds to ASR WAC using a single channel implementation of wavelet-based spectrogram reconstruction described in Section 2 and originally presented in [3]. For this system, the detection thresholds, \( \Omega \), were set using the Aurora 2 utterances from the speech babble noise condition at an SNR level 5 dB. The same \( \Omega \) values are used in the DWT filter bank of the single channel systems for all noise types. No results are displayed for the babble noise condition of the single channel system since that same data was used to determine the \( \Omega \) values. Finally, the last row of the noise type specific tables in Table 1, labeled “multi-channel”, corresponds
to the wavelet-based multi-channel fusion technique, described in Section 3.

There are several observations that can be made from Table 1. First, it is clear that all techniques provide significant WAC improvements with respect to the baseline for all conditions at low SNRs but less consistent improvements at high SNRs. Second, Table 1 shows that the single channel system for spectrogram reconstruction out-performs the MMSE system for all but the subway noise conditions. This poorer relative performance for the single channel system compared to the subway noise condition arises from both the mismatch in $\Omega$ settings for the single channel system as well as the fact that the MMSE system is particularly well suited to this class of stationary noise types. Finally, the last observation taken for Table 1 is that the wavelet-based multi-channel fusion approach consistently provides the highest WAC of all three approaches. The WAC for the multi-channel system represents an improvement over baseline performance for all but one high SNR condition.

Table 2 displays ASR WACs for the high-mismatch condition on the Spanish subset of the Aurora 3 corpus. The first column in Table 2 displays the baseline WAC, the second column displays the WAC for MMSE-based missing feature system, and the third column of Table 2 shows the WAC obtained for the wavelet-based multi-channel fusion system. The table shows an increase in WAC for the multi-channel system of 45 percent relative to the baseline and a 2.5 percent absolute increase in WAC with respect to a MMSE based system. This result is particularly important since it is obtained from speech collected in an actual noisy car environment.

5. Summary and Conclusions

An implementation of a closed-loop performance monitoring approach to combining multiple feature channels for robust ASR has been presented. The feature channels consist of spectrograms reconstructed from noise corrupted speech using multiple de-noising schemes in a multi-resolution DWT analysis-synthesis filter-bank. The approach has resulted in an increase in WAC of 45 percent relative to the baseline for the Aurora 3 speech in car noise ASR task. A 2.5 percent absolute increase in WAC was obtained with respect to a MMSE based missing feature approach to robust ASR on the same task. The effect of providing an ensemble of channels and a performance based criterion for channel selection appears to be an improved robustness with respect to varying noise types and SNRs. While the work and the experimental results reported here should be considered preliminary, it is believed that there is a great deal of potential in this general class of approaches. Future work will involve the investigation of more systematic approaches to creating richer ensembles of features channels as well as more comprehensive and efficient mechanisms for feature channel integration.

6. Acknowledgments

The authors would like to thank Hynek Hermansky at the Center for Language and Speech Processing at Johns Hopkins University for constructive advice and suggestions while Richard Rose served as a visiting researcher at the JHU Human Language Technology Center of Excellence.

7. References


