A COMPARISON OF PERFORMANCE MONITORING APPROACHES TO FUSING SPECTROGRAM CHANNELS IN SPEECH RECOGNITION

Shirin Badiezadegan and Richard Rose

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

ABSTRACT

Implementations of two performance monitoring approaches to feature channel integration in robust automatic speech recognition are presented. These approaches combine multiple feature channels, where the first one uses a feed-forward entropy-based criterion and the second one, motivated by psychophysical evidence in human speech perception, employs 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 these approaches is evaluated in the Aurora 3 speech in noise task domain.

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 speech spectrograms where each spectrogram corresponds to one of an ensemble of thresholding schemes. The parallel channels associated with these speech spectrograms are combined to generate ASR features through an entropy-based feed-forward and a feed-back "performance monitoring" approach to combining multiple channels. The feed-back performance monitoring approach is particularly 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. In the wavelet domain, the wavelet coefficients are "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 feed-back and feed-forward approaches described in Figure 1 and Figure 2 and presented in Section 3.

The overall approach presented here consists of three major steps . The first step is the extraction of parallel feature channels, $\boldsymbol{S} = \{S_1, \ldots, S_M\}$, from the corrupted speech. These channels correspond to the reconstructed spectrograms associated with the threshold settings $\boldsymbol{\Omega} = \{\Omega_1, \ldots, \Omega_M\}$ described above. The second step is to assess the quality of each feature channel, so that each channel is associated with a quality measure $\Phi_m, m = 1, \ldots, M$. As we see in Section 3, this step can be performed at frame level or at utterance level. Finally, the last step is to come up with a fusion strategy to either select or generate the spectrogram with the highest quality.

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 representing the original signal and suppresses 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, $\boldsymbol{y} = [y_1, \ldots, y_D]$, is extracted form noise corrupted speech. A speech presence probability (SPP) vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D]$ is estimated from a spectrographic mask. Each θ_d 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 $\boldsymbol{\theta}$ by propagating $\boldsymbol{\theta}$ through the DWT filter bank [3]. The mask propagation begins at the first wavelet scale where the mask vector, $\boldsymbol{\Theta}_1 = [\Theta_{1,1}, ..., \Theta_{1,K_1}]$, for the first scale is thresholded and applied to the wavelet coefficients $[Y_{1,1}, ..., Y_{1,K_1}]$ 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 $\boldsymbol{y}$, $[A_{1,1}, ..., A_{1,K_1}]$, resulting in mask components $\boldsymbol{\Delta}_1 = [\Delta_{1,1}, \ldots, \Delta_{1,K_1}]$ for the approximation coefficients. This process can be repeated so that the mask vectors at the jth scale, $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Delta}_j$, are propagated to the mask vectors at the $j+1$st scale, $\boldsymbol{\Theta}_{j+1}$ and $\boldsymbol{\Delta}_{j+1}$, up to the Jth 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, $\boldsymbol{Y}_j = [Y_{j,1}, ..., Y_{j,K_j}]$, 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 Λ_j:

$$\hat{\Theta}_{j,k} = \begin{cases} 1 & \Theta_{j,k} \geq \Lambda_j; \\ 0 & \Theta_{j,k} < \Lambda_j. \end{cases} \tag{1}$$

A similar approach is performed to map the continuous-valued elements of $\boldsymbol{\Delta}_j$ to binary-valued components of $\hat{\boldsymbol{\Delta}}_{\boldsymbol{j}}$ 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:

$$\hat{Y}^{hard}_{j,k} = \begin{cases} Y_{j,k} & \hat{\Theta}_{j,k} = 1; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

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, $\hat{\boldsymbol{y}}$.

3. CHANNEL FUSION

In this section, we take a closer look at the "channel fusion" block of the proposed systems depicted in Figure 1 and Figure 2. Specifically, we try to evaluate the "quality" of the reconstructed channels, $S_1, \ldots, S_M$. In Section 3.1, we show how we assess the quality of the feature channels at the frame level and exploit a feed-forward system to fuse the frames and to come up with the "best" enhanced spectrogram. Section 3.2 presents the utterance-level fusion scheme using a closed-loop system where the long-term quality of each utterance is measured and fed back to the fusion block to automatically select or generate the "best" enhanced spectrogram.

3.1. Feed-forward fusion

The DWT-based spectrogram reconstruction method processes a noisy utterance frame by frame. Thus, for a given noisy utterance with N frames and the DWT-based imputation system with M pre-defined thresholding settings, there exist $\binom{N\times M}{N}$ different ways to "re-assemble" the enhanced spectrogram. As depicted in Figure 1, we are exploiting an entropy-based measure to select the "best" enhanced frame out of M enhanced feature channels for each one of the N time frames and as a result, come up with the "best" spectrogram out of the $\binom{N\times M}{N}$ combinations. This is performed by estimating the entropy for each frame and exploiting that in a feed-forward system to automatically update the frames' fusion scheme.

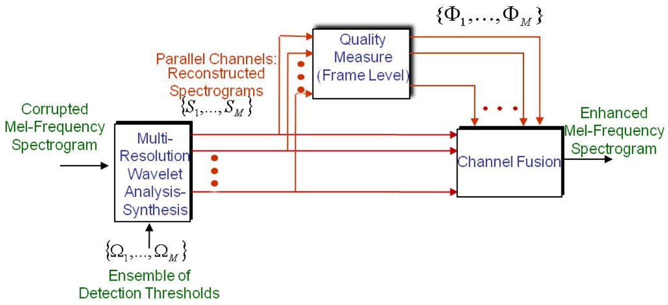
$\{\Phi_1,\ldots,\Phi_M\}$
Quality Measure (Frame Level)
Parallel Channels: Reconstructed Spectrograms
$\{S_1,\ldots,S_M\}$
Corrupted Mel-Frequency Spectrogram
Multi-Resolution Wavelet Analysis-Synthesis
Channel Fusion
Enhanced Mel-Frequency Spectrogram
$\{\Omega_1,\ldots,\Omega_M\}$
Ensemble of Detection Thresholds

Fig. 1: *Block diagram of the open-loop fusion system.*

To evaluate the entropy of the enhanced frames for each channel we first need to estimate the posterior probability of "sub-word" units using a single hidden layer multi-layer perceptron neural network (NN). This NN is trained from clean speech utterances to generate vectors of posterior probabilities, $\boldsymbol{P}_m,\ m = 1, \ldots, M$. 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. A vector consisting of the posterior probabilities at the output of the neural network is formed for each channel, $\boldsymbol{P}_m = [p_1, \ldots, p_L],\ m = 1, \ldots, M$, for a frame at a time. An entropy measure is assigned to each channel by:

$$\Phi_m = -\sum_{i=1}^{L} p_i\ \ln(p_i), \quad m = 1, \ldots, M. \tag{3}$$

It is believed that for a NN trained on clean speech, the entropy at the output of that NN increases in case of noisy speech [7]. Therefore, among the enhanced channels, the one with the minimum entropy is more likely to have the "highest" quality. With this criterion, the fusion scheme for the frame-level scenario becomes: $m^{opt} = \operatorname{argmin}\{\Phi_1, \ldots, \Phi_M\}$, and $\hat{\boldsymbol{y}}_n = S_{m^{opt}}$, in which $\hat{\boldsymbol{y}}_n$ is the "best" enhanced feature channel for frame n. By concatenating the $\hat{\boldsymbol{y}}_n$'s for all time frames $n = 1, \ldots, N$, the reconstructed spectrogram is obtained.

3.2. Feed-back fusion

This section describes the "channel fusion" portion of the system depicted in Figure 2. 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 observation vectors in Figure 2 correspond to posterior probabilities of observing sub-word units $p_l, l = 1, \ldots, L$ given the fusion of the input channels, $f_\alpha(\boldsymbol{S})$. Hence, component l of the observation vector $\vec{r}_n$ for frame n in an N frame utterance is given as $r_n^\alpha[l] = P(p_l|f_\alpha(\boldsymbol{S}))$. For the special case where α corresponds to a binary selection of input channels, this component of the observation vector is written as $r_n^{\alpha_m}[l] = P(p_l|S_{\alpha_m}[n])$ where $S_{\alpha_m}[n]$ is the vector of reconstructed spectrogram components for the nth frame of the mth channel.

To assess the quality of feature channel S_{α_m} for an N frame utterance, the first step is to estimate the sequence of posterior probability vectors $R^C_{\alpha_m} = [\vec{r}_1^{\alpha_m}, \ldots, \vec{r}_N^{\alpha_m}]$. The long-term statistics of this sequence are described by the autocorrelation matrix

$$C^C_{\alpha_m} = \sum_{n=1}^{N} \vec{r}_n^{\alpha_m}\ (\vec{r}_n^{\alpha_m})^T. \tag{4}$$

The accumulation of the $N \times N$ matrix, $C^C_{\alpha_m}$, 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 $C^C_{\alpha_m}$ 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 $R^C_{\alpha_m}$ 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 $C^C_{\alpha_m}$ deviates from the uncorrupted speech autocorrelation matrix, C^U. For corrupted utterances, the feature channels generating observations where $C^C_{\alpha_m}$ 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

$$\Phi_m = \sum_i \sum_j [C^C_{\alpha_m}]_{i,j} [C^U]_{i,j}, \tag{5}$$

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

The similarity measures $\Phi_1, \ldots, \Phi_M$ are fed back to the fusion block to determine the fusion parameters, α. In our implementations, we select the channel with the maximum Φ value, Φ_{max} if there is a significant difference between Φ_{max} and the next highest Φ. Otherwise, the optimum reconstructed spectrogram is computed from the weighted sum of the three channels with the highest Φ 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 Aurora 3 noisy

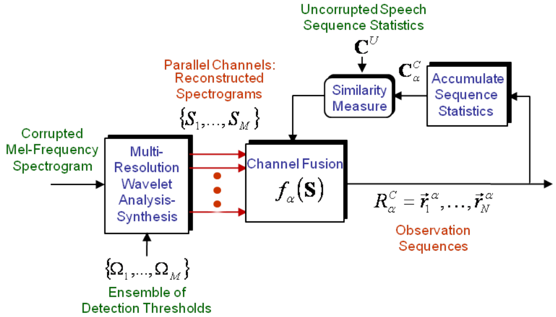

Fig. 2: *Block diagram of the closed-loop performance monitoring approach to channel fusion.*

speech task domain. The study will compare the performance of the multi-channel feed-back and feed-forward approaches. Moreover, the performance of the multi-channel systems is compared 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 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. The high mismatch condition [8] was 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.

The DWT based spectrogram reconstruction method discussed in Section 2 is implemented using a symlet 4 wavelet basis which has previously been used in a speech de-noising application [9]. 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 intuitive strategy is used for specifying a set of $M = 8$ channels of reconstructed spectrograms where the mth 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 Ω_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 $\{\Lambda_1^m, \ldots, \Lambda_j^m, \Gamma_1^m, \ldots, \Gamma_j^m\}$ are either "high" or "low" and the remaining threshold values which correspond to higher scales, $\{\Lambda_{j+1}^m, \ldots, \Lambda_4^m, \Gamma_{j+1}^m, \ldots, \Gamma_4^m\}$, are set to zero. This strategy limits the number of channels.

4.2. ASR performance

Table 1 displays ASR WACs for the high-mismatch condition on the Spanish subset of the Aurora 3 corpus. The first column displays the baseline WAC, the second column displays the WAC for MMSE-based missing feature system, and the third and forth columns show the WAC obtained for the wavelet-based multi-channel fusion, with the feed-forward and feed-back systems, respectively. The table shows an increase in WAC for the multi-channel systems of around 45 percent relative to the baseline and a 2 and 2.5 percent absolute increase in WAC with respect to a MMSE based system for the feed-forward and feed-back systems, respectively. This result is particularly important since it is obtained from speech collected in an actual noisy car environment. The other interesting observation is that the feed-back system outperforms the feed-forward approach. This observation implies the importance of considering the long-term statistics of the channels (as implemented in the feed-back system) in channel fusion.

5. SUMMARY AND CONCLUSIONS

Implementations of a feed-forward and a feed-back performance monitoring approach to combining multiple feature channels for robust ASR have been presented. A 2.5 percent absolute increase in WAC was obtained for the feed-back system with respect to a MMSE based missing feature approach to robust ASR on the same task. The feed-forward system also resulted in an increase in WAC with respect to the MMSE approach, but was outperformed by the feed-back system.

The feed-back system provides a mechanism for determining when the long-term statistical behavior of the system departs from the statistics of the uncorrupted system. The feed-forward system, on the other hand, measures the quality of the short-term statistics of the system. This advantage of the feed-back systems explains the higher WAC obtained for the closed-loop system compared to the feed-forward approach.

Table 1: *ASR WAC for DWT-based imputation on Aurora3, Spanish, high mismatched subset.*

Method	Baseline	MMSE	Feed-forward	Feed-back
WAC(%)	48.72	68.54	70.50	71.01

6. REFERENCES

[1] Mesgarani, N., Thomas, S., and Hermansky, H., "Toward optimizing stream fusion in multistream recognition of speech", submitted to J. Acoust. Soc. Am. Express Letters, March 2011.

[2] Scheffers, M.K. , and Coles, M.G., "Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors", Journal of Experimental Psychology: Human Perception and Performance, pp. 141-151, vol. 26, no.1, 2000.

[3] Badiezadegan, S. and Rose, R. C., "A wavelet-based data imputation approach to spectrogram reconstruction for robust speech recognition ", in Proc. ICASSP, Czech Republic, 2011.

[4] Dohoho, D.L. and Johnstone, I.M.,"Ideal spatial adaptation by wavelet shrinkage", Biometrika, pp. 425-455, vol. 81, no. 3, 1994.

[5] Raj, B. and Singh, R., "Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition", in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 65-70, 2005.

[6] Badiezadegan, S. and Rose, R. C., "Mask estimation in non-stationary noise environments for missing feature based robust speech recognition", in Proc. Interspeech, Japan, 2010.

[7] Misra, H., Bourland, H. and Tyagi, V.,"New entropy based combination rules in HMM/ANN multi-stream ASR", in Proc.

[8] Badiezadegan, S. and Rose, R., "A Performance Monitoring Approach to Fusing Enhanced Spectrogram Channels in Robust Speech Recognition," submitted to Interspeech 2011, Italy.

[9] Soon, I.Y. and Koh, S.N. and Yeo, C.K.,"Wavelet for speech denoising", in Proc. IEEE Region 10 Annula Conference on Speech and Image Technologies for Computing and Telecommunications, TENCON-97, pp. 479-482, vol.2, 1997.