A Noise-type and Level-dependent MPO-based Speech Enhancement Architecture with Variable Frame Analysis for Noise-robust Speech Recognition

Vikramjit Mitra[1], Bengt J. Borgstrom[2], Carol Y. Espy-Wilson[1], Abeer Alwan[2]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD
[2]Department of Electrical Engineering, University California, Los Angeles, CA
[1]{vmitra@umd.edu, espy@umd.edu}, [2]{jonas@ee.ucla.edu, alwan@ee.ucla.edu}

Abstract

In previous work, a speech enhancement algorithm based on phase opponency and a periodicity measure (MPO-APP) was developed for speech recognition. Axiomatic thresholds were used in the MPO-APP regardless of the signal-to-noise ratio (SNR) of the corrupted speech or any characterization of the noise. The current work developed an algorithm for adjusting the threshold in the MPO-APP based on the SNR and whether the speech signal is clean, corrupted by aperiodic noise or corrupted with noise with periodic components. In addition, variable frame rate (VFR) analysis has been incorporated so that dynamic regions in the speech signal are more heavily sampled than steady-state regions. The result is a 2-stage algorithm that gives superior performance to the previous MPO-APP, and to several other state-of-the-art speech enhancement algorithms.

Index Terms: Speech enhancement, robust speech recognition, SNR estimation, variable frame rate analysis, phase opponency.

1. Introduction

Several approaches have been explored to improve noise-robustness of automatic speech recognition (ASR) systems. One such approach is to enhance the speech signal by suppressing the noise while retaining the speech contents undistorted. Such a technique is usually used within the front-end of an ASR system prior to estimating the features that will be used for recognition. The Modified Phase Opponency (MPO) algorithm together with a periodicity summary measure (MPO-APP) has been proposed [1] as a stand-alone speech enhancer for this kind of application. The MPO-APP passes as-is the harmonics in the formant regions of the speech signal and attenuates the rest. Objective measures show that the MPO-APP has comparable performance to many other speech enhancement techniques, but has superior performance in fluctuating noise [1]. However, at low SNRs (below 0 dB), MPO-APP is found to suffer from speech deletions. This happens due to the severe masking effect of the noise, especially at higher frequencies, whose energy is usually weaker due to the natural roll-off of the glottal spectrum.

Recent studies on the functional dependence of the periodicity summary thresholds (θ_{th}) on ASR recognition accuracy for the MPO-APP enhanced speech reveals that for optimal performance, θ_{th} should vary with the SNR and also with the noise type. Thus, optimal performance necessitates *a priori* knowledge about the noise type and it's SNR for the input speech, which are usually unknown for real-world problems. The studies also showed that at a given SNR, the thresholds for broad categories of the noise (i.e., aperiodic noise, noise with periodic components and clean speech) show similar values. Also, for a given noise type, the thresholds are found to vary almost linearly with SNR levels. These two facts suggest that given *a priori* knowledge about the broad noise type and its SNR, sub-optimal θ_{th} values can be predicted which ensures near-optimal performance of the MPO-APP in terms of ASR accuracy. It is also observed that the performance of the MPO-APP enhancement improves with an increase in the SNR.

These observations led us to a 2-stage architecture as a refinement to the MPO-APP algorithm. In the first stage, a preprocessor is used to (a) estimate the SNR of the input speech, (b) reduce the quasi-stationary noise component from the input speech, (c) obtain the broad noise type of the input speech (i.e., whether the noise is clean, or contains a periodic component (e.g., babble noise, airport noise etc which typically has a multitude of background speakers), or is strictly aperiodic in nature (e.g., car noise, subway noise etc.), and (d) predict near optimal periodicity thresholds for MPO-APP based upon the estimated SNR and the detected broad noise type. A more aggressive noise reduction is performed in the second stage using the MPO-APP architecture, which uses the optimal periodicity thresholds obtained from the first stage. Prior to recognition, feature extraction is performed using variable frame rate (VFR) analysis [8, 9]. VFR analysis aims to oversample frames during speech segments showing high spectral dynamics, which can be considered important for recognition, while sparsely sampling steady-state segments. Results indicate that the combination of the 2-stage MPO-APP based speech enhancement architecture with VFR provides considerable improvement in recognition performance in noisy conditions over the baseline as well as some of the other noise-robust approaches.

2. The data

The evaluation of the preprocessor based MPOAPP with VFR is performed by using ASR experiments on the Aurora-2 [3] dataset. The Aurora-2 dataset is created from the TIdigits dataset, which consists of connected digits spoken by an American English talker sampled at 8kHz. There are three sections in this data set, test-set A, B and C; where sets A and B have four subparts representing four different noises (altogether eight different noise types). Section C contains two subsections representing two noise types from section A and B, but involving a different channel. As channel effects are not considered in our approach, test-set C is ignored in this research. Only the training in clean and the testing in noisy scenario was used.

3. The Pre-processor based MPO-APP

The MPO-APP [2, 4] first performs MPO enhancement on the speech and then a summary periodicity confidence measure [2] is used to remove noise insertions (occurs when

the noise has narrow-band resonances) and reinsert speech deletions (occurs when the speech signal appears wide band due to two closely-spaced formants, e.g., when F2 and F3 moves closer to each other in /r/). The present work uses a summary periodicity energy measure ($\hat{E}_{per}$) as opposed to a summary periodicity confidence measure to obtain θ_{low} and θ_{high}, as the former is found to be the more reliable indicator of voicing.

The periodicity measure uses two thresholds. The lower threshold, θ_{low}, deals with noise insertions and a higher threshold, θ_{high}, deals with speech deletions. In the initial implementation of MPO-APP [4], θ_{low} was estimated from clean speech and it was assumed that θ_{low} does not change with varying background conditions. Keeping θ_{high} equal to θ_{low}+2, we varied θ_{low} and performed speech enhancement on a subset of the Aurora corpus that we call the "dev-set" (200 files selected randomly from each noise type and SNR level) and the recognition results are shown in Figure 1. It shows that (a) the optimal ASR performance depends upon the careful selection of θ_{low} (which we term as θ_{lowopt}) and (b) θ_{lowopt} varies with SNR.

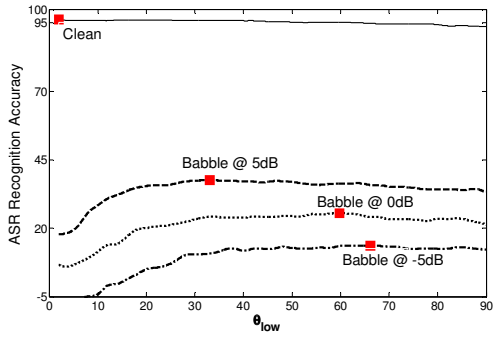

Figure 1. Plot of ASR accuracy on the dev-set obtained by varying θ_{low}, and maintaining θ_{high}=2+θ_{low}, for clean speech and speech with babble noise at 5dB, 0dB and -5dB SNRs. The square points represent the Rec. acc. at θ_{lowopt} for babble at that specific SNR.

Figure 2 shows the effect of θ_{low} on the ASR accuracy across different noise types at 20dB SNR for the dev-set. Figure 2 presents the following interesting observations: (a) at a specific SNR, the θ_{lowopt} is found to vary with noise type and (b) θ_{lowopt} for aperiodic noise types (car, street, train-station, subway and exhibition noise; where recognition accuracy at θ_{lowopt} is shown in solid squares) show similar values and so does θ_{lowopt} for noises having periodic components due to background speakers (babble, airport, and restaurant noise; where recognition accuracy at θ_{lowopt} is shown in solid circles). The θ_{lowopt}'s for these two broad noise types are very different from one another. *A priori* knowledge about the noise types may not be possible for real world applications; however, broad noise types can be detected and hence suboptimal performance of the MPOAPP can be ensured.

In the first stage of the proposed preprocessor based MPO-APP architecture, a VAD is constructed, whose decision is based upon $\hat{E}_{per}$ and other parameters obtained from the APP detector [5]. Figure 3 shows $\hat{E}_{per}$ for a speech file from Aurora which has been corrupted with subway noise at 10dB SNR. A threshold is used (which is a function of the statistics and the dynamic range of $\hat{E}_{per}$, details about this procedure are beyond the scope of this paper) to distinguish between speech-dominant and speech-absent frames. The knowledge of the speech-dominant and speech-absent frames is used for: (a) estimating the SNR of the input speech, (b) obtaining the noise spectrum from speech absent frames to perform initial speech enhancement using signal theoretic approaches and (c) identifying the broad noise type from the spectral information of speech absent frames.

3.1. Estimating the SNR

The signal power at the speech-dominant frames and speech-absent frames was used to estimate the SNR. Table 1, presents the SNR estimates averaged across test sets *A* and *B*.

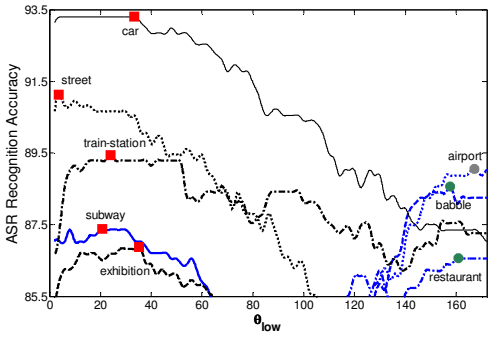

Figure 2. Plot of ASR accuracy on the dev-set obtained by varying θ_{low}, and maintaining θ_{high}=2+ θ_{low}, for all noise types in Aurora at 20dB SNR. The solid points represent the θ_{lowopt} for each noise type.

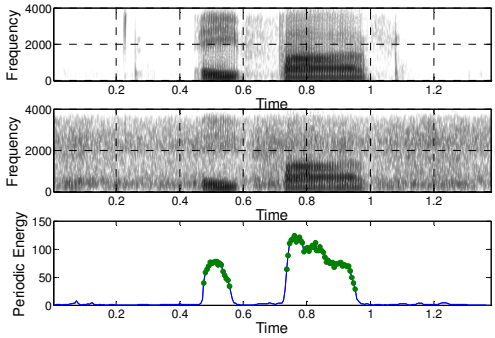

Figure 3. Plot showing the performance of the VAD, the topmost spectrogram shows the clean speech ("eight five"), the second spectrogram shows the signal corrupted with subway noise at 10dB, the curve at the bottom shows $\hat{E}_{per}$, where the thicker regions are detected as speech present.

Table. 1. SNR Estimate (Average for Aurora-2 test-set *A* & *B*)

		Prior SNR Estimate	
		Mean (dB)	Standard dev
Actual SNR	20dB	17.78	1.78
	15dB	13.01	1.64
	10dB	8.22	1.69
	5dB	3.41	1.66
	0dB	-0.65	1.74
	-5dB	-2.82	1.79

3.2. Pre-processing by signal theoretic approaches to lower quasi-stationary noise

The MPO-APP passes narrow band components as-is without attenuating them. As a result, noise gets passed along with the speech in the speech-dominant regions (i.e., formant regions). When the speech signal is not sufficiently dominant to mask the noise, the noise becomes perceivable creating a 'shadow-effect'. It was also observed that the performance of the MPO-APP improves with increase in SNR. To address these issues, the knowledge of the speech-absent frames is exploited

to perform initial enhancement of speech using signal theoretic approaches to reduce the quasi-stationary component of the noise. Two approaches have been considered for the initial enhancement: generalized spectral subtraction (GSS) [6] and speech enhancement by minimum mean square log-spectral amplitude estimator (LMMSE) [7]. The knowledge of the speech absent frames is used to obtain a local average noise spectrum using equation (1)

$$\hat{S}_{n,t}(\omega) = \alpha \hat{S}_{avg}(\omega) + (1-\alpha)\hat{S}_{lavg,t}(\omega) \quad if \quad \hat{S}_{lavg,t}(\omega) \neq 0$$
$$\hat{S}_{n,t}(\omega) = \hat{S}_{avg}(\omega) \quad otherwise \qquad (1)$$

where $\hat{S}_{n,t}(\omega)$ is the noise spectrum estimate for a particular frame t, $\hat{S}_{avg}(\omega)$ is the average noise spectrum over all the frames, $\hat{S}_{lavg,t}(\omega)$ is the local average noise spectrum for frame t considering 10 frames before and 10 frames after t, and α is considered to be 0.5.

3.3. Estimating the broad category of the noise

The speech-absent frames were used to identify the broad-noise type. Thirteen Linear Prediction Cepstral Coefficients (LPCC) are generated for the speech absent frames, using an analysis window of 20ms and an advance of 5ms, where the coefficients are normalized by C_0, yielding 12 coefficients. These coefficients after normalization were used to train a 2-hidden layer feed-forward Artificial Neural Network (ANN) with a goal to discern between clean speech, speech with aperiodic noise and speech with periodic noise components. It was empirically found that hidden layers with 75 processing elements (PE) in the first layer and 50 PEs in the second layer (with biases and tan-sigmoid transfer function in all layers) generated optimal results in terms of accuracy as well as computation time. Back-propagation with scaled conjugate gradient training was used to train the ANN, using the dev-set as the training corpus. The recognition accuracy for all the files in test-set A and B was 94.1%.

The preprocessor based MPO-APP architecture is shown in Figure 4. θ_{low} for clean speech is globally set to be 14.0, which is inferred from empirical results. For the aperiodic noise or the noise with a periodic component, the appropriate piecewise linear curve shown in Figure 5 is used to obtain θ_{low} given the SNR estimate. Figure 5 is different from Figure 1 in the sense that the former uses the estimated mean SNR where as the latter uses the SNR specified by Aurora database. Both GSS and LMMSE suffer from deletion errors at low SNR. Feeding the enhanced speech from either of these two algorithms to MPO-APP ensures propagation of deletion errors. To circumvent this, a mixing of the original noisy speech is performed with the GSS or LMMSE enhanced speech, which ensures an increase in the SNR at the same time the noisy component reduces the probability of deletion error by MPO-APP, even if LMMSE and GSS suffers from those errors (see Figure 6 which shows that GSS deleted the 2nd "zero", however, the mixing allowed the MPO to detect that region). The mixing is performed using equation 2

$$S_{mix}[n] = \beta S_{pre}[n] + (1-\beta)S[n] \quad where, \ 0 \le \beta \le 1 \quad (2)$$

where $S_{mix}[n]$ is the mixed signal which is fed to the MPO-APP, β is the mixing coefficient, $S_{pre}[n]$ is the pre-enhanced speech and $S[n]$ is the noisy speech. ASR experiments at different noise types and noise levels suggest that β is a function of SNR and noise type. This is because (1) MPO-APP is known to perform well at high SNRs, hence at higher SNRs, β can be close to zero if not zero, (2) as the SNR is lowered, performance of the MPO-APP degrades, hence, improving the SNR by increasing β helps to improve the performance of MPO-APP but β is constrained by the amount of deletion errors suffered by GSS or LMMSE. The estimated SNR and the noise broad-class are used to obtain the lower periodicity threshold, θ_{low} and the mixing coefficient, β (where β is also modeled as a piecewise linear function of SNR (based on empirical data), given broad noise type). ASR experiments with θ_{high} shows that it has relatively less contribution to ASR accuracy as compared to θ_{low}. Hence no direct estimation of θ_{high} was performed. Instead, it is set to $\theta_{high} = 1.46 + \theta_{low}$. In the second stage the thresholds and the partially cleaned signal from the first stage is used by the MPO-APP to perform a more aggressive speech enhancement.

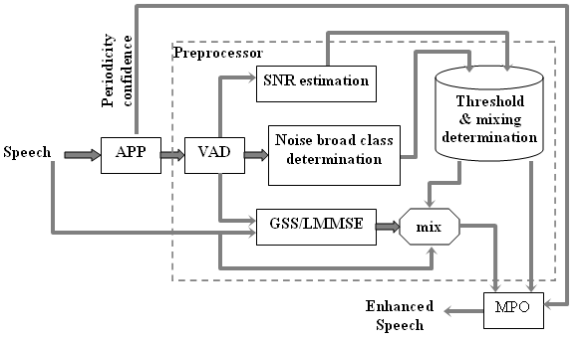

Figure 4. Block-diagram of the preprocessor based MPO-APP

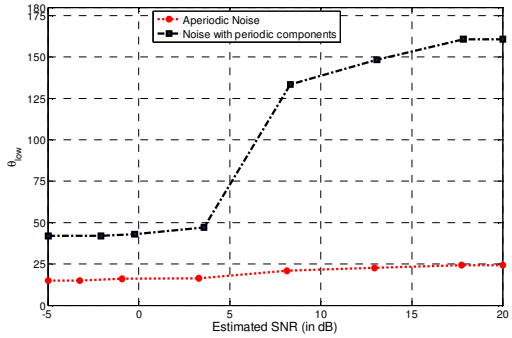

Figure 5. Curves used to estimate θ_{low} given the estimated SNR and the knowledge of the broad noise type.

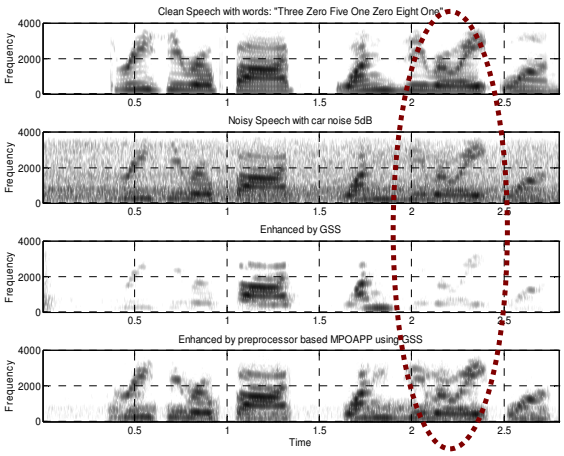

Figure 6. Spectrograms showing (a) clean speech with words "Three Zero Five One Zero Eight One", (b) the same speech with 5dB car noise, (c) the noisy speech after enhanced by GSS (d) the noisy speech after MPO-APP enhancement using GSS

4. Variable Frame Rate (VFR) analysis

Variable frame rate (VFR) analysis aims to oversample frames during speech segments showing high spectral dynamics, since these regions have been shown to be perceptually important, while sparsely sampling other segments. Thus, perceptually rich signal segments are captured at higher resolutions. VFR has been shown to improve the noise robustness of ASR [8], [9]. An integral component of VFR analysis is VAD since it allows for sparse sampling of frames during non-speech segments, even if the corresponding signal shows relatively high spectral dynamics. However, since the performance of VAD algorithms tend to degrade at lower SNRs, the overall performance of the VFR feature extraction system can be expected to degrade as well. The high noise suppression during non-speech segments achieved by the proposed enhancement architecture makes the VAD a simpler task, and even allows for the use of less complex VAD algorithms. In this study, VFR analysis was implemented using a low-complexity VAD based on frame energy.

5. Results

Results show that the GSS based architecture offers better recognition accuracy than its LMMSE counterpart. The feature vector and the HMM model for the experiments in our research are kept the same as specified for use with the Aurora-2 dataset [3]. Figure 7 shows the average ASR recognition accuracy for only aperiodic noise types whereas Figure 8 shows the same for noises having periodic components. From Figures 7-8 it is evident that the preprocessor based MPO-APP outperformed the MPO-APP proposed in [2]. Table 2 presents the word accuracies according to Hirsch and Pierce [3], where the word accuracies

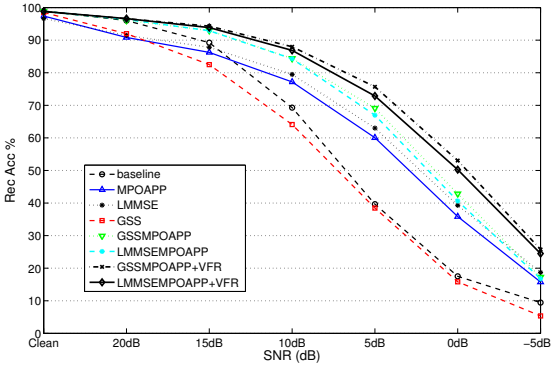

Figure 7. Average recognition accuracy for aperiodic noise types

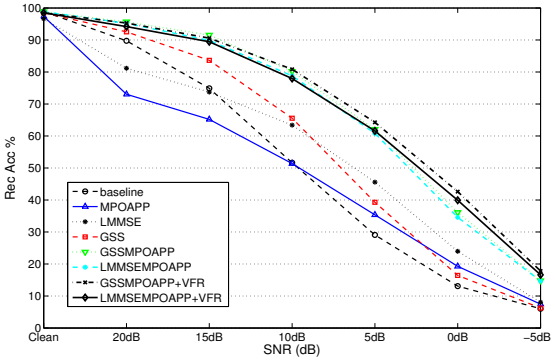

Figure 8. Average recognition accuracy for periodic noise types

for a test set are averaged across all the noise types for SNRs between 0dB and 20dB. The table suggests that the preprocessor based MPOAPP enhancement helped to improve the recognition accuracy significantly over the baseline and the performance is improved further by the VFR analysis.

Table 2. Average word recognition accuracy for the Aurora-2 database

	Absolute word accuracy %	
	Set A	Set B
Baseline	61.13	55.57
LMMSEMPOAPP	74.34	74.83
LMMSEMPOAPP+VFR	77.61	76.98
GSSMPOAPP	75.13	76.05
GSSMPOAPP+VFR	79.09	79.13

6. Conclusion

Comparing the contents of Table 2, it can be seen that GSSMPOAPP+VFR offered the best recognition accuracy. Also from Figure 7-8, it is evident that the preprocessor based MPO-APP architecture consistently outperformed the other competing enhancement schemes considered in this study. MPOAPP uses a spectro-temporal mask to discern noise dominant regions from speech dominant ones; future research should consider using this mask to implement a missing feature technique to further improve ASR noise robustness. Currently MPOAPP attenuates the noise with a constant factor, future research should exploit the estimated SNR to adaptively control the attenuation, which would help to retain consonantal information at higher SNRs and may improving ASR accuracy at those SNRs.

Acknowledgements

This research was supported by NSF Grant # IIS0703859.

7. References

[1] O. Deshmukh and C. Espy-Wilson, "Modified Phase Opponency Based Solution to the Speech Separation Challenge", In Proc. of Interspeech 2006, pp. 101-104, Pittsburgh, PA.

[2] O. Deshmukh, C. Espy-Wilson and L.H. Carney, "Speech Enhancement Using The Modified Phase Opponency Model", Journal of Acoustic Society of America, Vol. 121, No. 6, pp 3886-3898, 2007.

[3] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", In Proc. ISCA ITRW ASR2000, pp. 181-188, Paris, France, 2000.

[4] L.H. Carney, M.G. Heinz, M.E. Evilsizer, R.H. Gilkey and H.S. Colburn, "Auditory phase opponency: A temporal model for masked detection at low frequencies", Acustica - Acta Acustica, Vol. 88, pp. 334-347, 2002.

[5] O. Deshmukh, C. Espy-Wilson, A. Salomon and J. Singh, "Use of Temporal Information: Detection of the Periodicity and Aperiodicity Profile of Speech", IEEE Transactions on Speech and Audio Processing, Vol. 13(5), pp. 776-786, 2005.

[6] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. Speech Audio Process. Vol.7, No.2, pp. 126-137, 1999.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square log-spectral amplitude estimator", IEEE Transactions on Acoustics Speech and Signal Processing, Vol. ASSP-33(2), pp. 443–445, 1985.

[8] H. You, Q. Zhu, and A. Alwan, "Entropy-Based Variable Frame Rate Analysis of Speech Signals and its Application to ASR", ICASSP, pp. 549-552, 2004.

[9] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition", ICASSP, pp. 3264-3267, 2000.