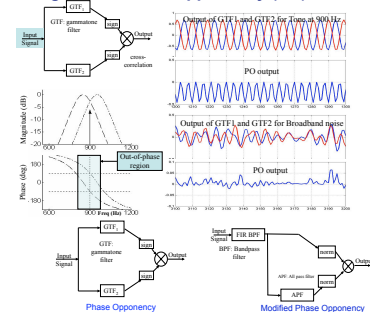


Introduction

In previous work, we developed a speech enhancement algorithm based on Phase Opponency [1] called the Modified Phase Opponency (MPO) [2] model. In the present work, we discuss extensions of the model to further improve its performance.

- The extension proposes a preprocessor, which
 - Performs Voice activity detection (VAD) on the input speech signal
 - Initially reduces the quasi-stationary component of the noise
 - Detects the SNR of the input speech signal
 - Ensures near-optimal performance of the MPO-based speech enhancement

Background: Phase Opponency (PO)



APP provides a periodicity summary measure, which is the frame-wise total periodic energy in a signal

- Two thresholds (θ_{low} and θ_{high}) are used to deal with noise insertions and speech deletion issues with MPO

Issues with MPO-APP

- The periodicity threshold used to distinguish narrow-band noise regions and wide band speech regions was kept the same for all noise types and levels even though empirical studies show that it varies as a function of both (see Fig. 1 where we varied them in steps and performed speech recognition experiments on the enhanced speech).

Fig. 1 Plot of ASR accuracy by varying θ_{low} and maintaining $\theta_{high}=2*\theta_{low}$ for clean speech and speech corrupted with babble noise at 5dB, 0dB and -5dB SNRs. The red square points represent the point where maximal accuracy is obtained

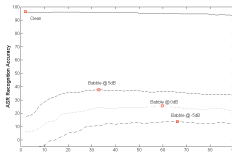
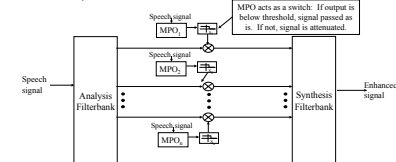


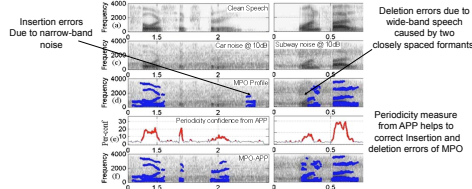
Fig. 2 Plot of ASR accuracy by varying θ_{low} and maintaining $\theta_{high}=2*\theta_{low}$ for all noise types in Aurora-2 at 20dB SNR. The solid square points represent the point where maximal accuracy was obtained

- Same set of parameters control relative magnitude responses and relative phase responses.
- Difficult to predict the relation between the location and the width of the out-of-phase region and the filter parameters.
- Separate parameters control relative magnitude and phase responses \rightarrow can manipulate them independently.
- APF \rightarrow only phase difference.
- BPF \rightarrow only freq. selectivity.
- Easy to analyze and control.

MPO based speech enhancement



Problems with MPO and how MPO with APP [3,4] addresses them



Figures 1 & 2 show

- Optimal ASR performance depends upon the careful selection of θ_{low}
- Optimal θ_{low} varies with SNR and with noise-type.
- For aperiodic noise types (e.g. car, subway, train-station etc.) the optimal θ_{low} thresholds have similar values (tight bound).
- For noise with periodic components (e.g. babble, airport etc.), the optimal θ_{low} thresholds have similar values (tight bound).
- Optimal θ_{low} are largely different for aperiodic noise types and periodic noise types.

A priori knowledge about noise type is usually unknown, but broad noise types (i.e. whether a noise has predominantly aperiodic or periodic components) can be recognized.

Role of the Preprocessor

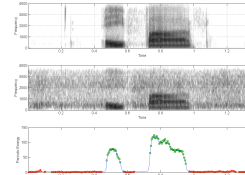
- Obtain an initial estimate of the SNR
- Obtain the broad-noise type (i.e. noise with aperiodic component, noise with periodic component and clean speech)
- Estimate near optimal thresholds θ_{low} and θ_{high} based on SNR and broad noise-type.

- Reduce the noise (increasing SNR) before MPO-APP processing

- MPO passes the signal (speech+noise) as-is in the speech-dominant regions. At low SNRs this result in passing considerable noise between the speech harmonics so that the speech fails to mask the background noise.
- Overall MPO-APP is found to perform well if SNR is high

- To estimate the SNR: A Voice Activity detector (VAD) based on the APP detector distinguishes the speech-dominant frames from the speech-absent frames. The SNR is estimated from the signal power and the noise power from the speech-present and speech-absent frames, respectively.
- To estimate the broad noise type: Cepstral coefficients from the speech-absent frames are used by an artificial neural network (ANN) to obtain the broad noise type for the input speech.
- Speech-absent frames are used by signal theoretic enhancement schemes (generalized spectral subtraction (GSS) and speech enhancement by minimum mean square log-spectral amplitude estimator (LMMSE)) to perform initial cleaning of noise.

Fig. 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 periodicity summary measure from APP, where the red starred regions are detected as speech absent and the green circled regions are detected as speech present



Block-diagram of the preprocessor based MPO-APP

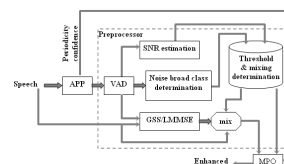
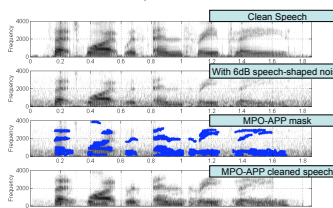


Illustration of the performance of MPOAPP



Database

- Evaluation of the proposed front end was performed using the Aurora-2 dataset
- Only the training in clean and the testing in noise was used
- Test sections A & B were considered, which contain 8 different noise types.
- Test section C was not considered as it dealt with channel differences.

Evaluation

The enhanced speech was used to generate 13 Mel-frequency cepstral coefficients (MFCC) along with their delta and acceleration coefficients, generating a 39-dimensional feature set.

Variable frame rate (VFR) [5] has been incorporated so that dynamic regions in the speech signal are more heavily sampled than steady-state regions

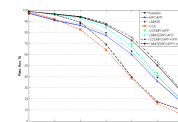
16-state, 6-mixture word models were used in the recognizer

Results

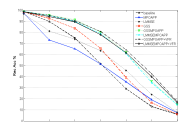
SNR Estimate (Average for Aurora-2 test-set A & B)

		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

Average recognition accuracy for aperiodic noise types



Average recognition accuracy for periodic noise types



Conclusion

- The proposed preprocessor with MPOAPP helped to improve ASR noise robustness
- VFR along with GSS-MPOAPP offered the best result

Future Directions

- MPOAPP attenuates the noise with a constant factor at present
 - Attenuation can be made as a function of SNR estimate
 - Lower attenuation at high SNRs can help to retain consonantal information, which can improve ASR accuracy at those SNRs.
- MPOAPP/GSS uses a spectro-temporal mask to discern noise dominant regions from speech dominant ones. This mask can be used for missing feature technique to further improve ASR noise robustness.

References

- L. Carney, M. D. Heinz, M. E. Evlizler, R. L. Gheib, H. S. Colburn, "Auditory phase opponency: A temporal model for masked detection at low frequencies", *Acta Acust.* 88, 334-347, 2002.
- O. D. Deshmukh and C. Espy-Wilson, "Modified Phase Opponency Based Solution to the Speech Separation Challenge", In *Proc. of Interspeech 2005*, pp. 101-104, Pittsburgh, PA.
- O. D. 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.
- O. D. 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.
- G. Qiu and A. Alwan, "On the use of variable frame rate analysis in speech recognition", *ICASSP*, pp. 3284-3287, 2000.

Acknowledgement

This research was supported by NSF grant IIS0703859.