Robust Feature Compensation in Nonstationary and Multiple Noise Environments

Martin Graciarena, Horacio Franco, Greg Myers and Victor Abrash

Speech Technology and Research Laboratory,
SRI International, USA
martin@speech.sri.com

Abstract

The probabilistic optimum filtering (POF) algorithm is a piece wise linear transformation of the noisy speech feature space into the clean speech feature space. In this work we extend the POF algorithm to allow a more accurate way to select noisy-to-clean feature mappings, by allowing different combinations of speech and noise to have combination-specific mappings selected depending on the observation. This is especially important in nonstationary environments, where different noise segments will result in different observations in the noisy feature space. Experimental results using stationary and nonstationary noises show the effectiveness of the proposed technique compared to the old approach. We also explored the use of the extended POF method to train a map with multiple noises in order to gain generalization over different noise types and be able to tackle unknown noise environments.

1. Introduction

Speech recognition technology has the potential to increase interaction efficiency from human-machine speech interfaces for command and control and information access to speech-to-speech translation systems. However, the high levels of noise often found in the real-world applications present a significant challenge to speech recognition systems. Without specific noise-robust processing, even state-of-the-art speech recognition degrades rapidly under decreasing signal-to-noise ratios (SNR). Even when noise-canceling microphones are used, the speech signal may have levels of noise that can impact accuracy significantly, especially in typical field conditions and military vehicles [1].

One of the many noise-robust processing techniques is feature compensation. This technique aims at reducing the mismatch between the testing and training conditions by producing an estimate of the clean speech features from the noisy speech features. Several feature compensation algorithms have been proposed; among them we can cite the Probabilistic Optimum Filtering (POF) algorithm [2], the SPLICE algorithm [3], and the MEMLIN algorithm [4].

In this work we extend the POF algorithm to allow a more accurate way to select noisy-to-clean feature mappings. Our goal is to allow different combinations of speech and noise to have speech and noise combination-specific mappings selected depending on the observed noisy speech. This is especially important in nonstationary environments, where different noise segments when combined with the clean speech will produce different observations in the noisy feature space. We present experimental results that show the advantage of the proposed extension. We also explore the use of the extended POF approach to train a map with multiple noises in order to gain generalization over different noise types and be able to tackle unknown noise environments.

2. The Probabilistic Optimum Filtering Algorithm

The POF algorithm [2], is a piece wise linear transformation of a noisy feature space into a clean feature space. Each linear transformation is assigned to a region in a vector quantization (VQ) partition of the feature space. The POF mappings are obtained from training samples of clean speech and of noise obtained from the target environment. To train the mappings we randomly add the collected noise samples to the clean speech at different SNRs; then we derive synchronized feature streams from both the clean and the noisy versions of the speech data. We use those two sets of feature streams to create the POF mappings for each SNR. These two sets are what we call a "stereo" database. These POF mappings can be trained with a moderate amount of data, and are simpler and faster to develop than new acoustic models trained for each noisy acoustic environment. During recognition, different POF mapping sets can be dynamically selected based on real-time estimates of the SNR of the current condition, to find the closest match between training and test conditions.

More formally, if $\hat{x}_t$ is the POF-derived clean speech feature vector at frame index *t*, Y_t is the tapped-delay line of noisy feature vectors (with n noisy vectors, where n is the window parameter), I is the number of VQ regions, z_t is the conditioning vector, which in our case is the noisy speech feature, and W_i is the filter coefficient matrix associated with region *i*, then the POF estimate can be computed as follows $\hat{x}_t = \sum_{i=0}^{I-1} \{W_i^T p(g_i / z_t)\} Y_t$, where $p(.)$ denotes the probability that the clean vector x_t belongs to the VQ region g_i given the noisy conditioning vector z_t. The W_i matrices are trained so as to minimize the squared prediction error in the "stereo" database. The probability can be computed using the Bayes rule as follows $p(g_i / z_t) = p(z_t / g_i) p(g_i) / p(z_t)$, where $p(z_t / g_i)$ is the likelihood of the noisy feature given the VQ region g_i, $p(g_i)$ is the prior probability of the VQ region g_i, and $p(z_t)$ is the probability of the noisy speech feature.

3. New POF Approach

In the original POF algorithm the VQ regions were defined in the clean feature space. They were computed using the Lloyd VQ estimation algorithm [5]. From each clean speech feature in a VQ region and using the stereo database, the corresponding noisy speech feature can be known. From the cluster of the noisy speech features, corresponding to the same VQ region, a Gaussian probability density was computed.

The problem with this approach is that in the case of nonstationary noise the noisy speech feature observations may be more spread than in the stationary noise case. Therefore, the resulting Gaussian density may have a high variance and there might be severe class overlap.

In Figure 1 we present a simple case of the original approach of VQ clustering in the clean feature space for stationary noise. Assuming three different clean speech clusters in the clean speech feature domain, since the noise is stationary this will correspond roughly to three noisy speech feature domain clusters. The Gaussian densities computed in the noisy speech feature domain are well defined and the classes are well separated.

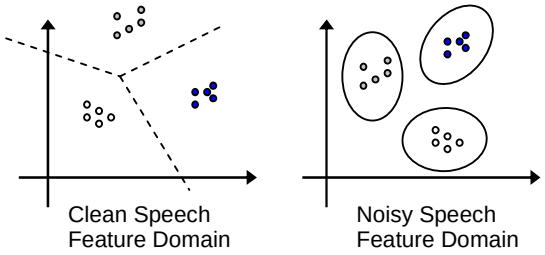

Figure 1: VQ Clustering in Clean Speech Feature Domain in Stationary Noise

In Figure 2 we explore the same clean speech feature domain clusters but with a nonstationary noise. For the sake of simplicity we assume a nonstationary noise to be composed by a sequence of two different noises. Therefore, for each clean speech cluster we have two different noisy speech feature domain clusters. The Gaussian densities computed for each cluster in the noisy space now have a high variance and the classes are not well defined. This will also result in poor training of the linear transformations.

Note that increasing the number of VQ clusters in the clean feature domain and therefore the number of Gaussians in the noisy feature domain cannot solve this problem. For each new VQ cluster we may have the same problem of multiple observations in the noisy speech feature domain.

To have a more detailed modeling of the noisy space we could assign a Gaussian mixture in the noisy speech feature domain to each VQ cluster in the clean speech feature domain. However, this may result in averaging clean speech feature estimates, via mixture weights, from different clusters of noisy speech features.

A more direct approach is to compute the VQ partitioning in the noisy feature acoustic space instead. This way of partitioning allows us to have different mappings corresponding to the same speech feature when it is combined with different noise types. Thus, speech and noise combination-specific feature mappings can be defined and trained. Speech and noise combination-specific mappings should produce more accurate estimates of the clean features when several types of noises or varying noisy conditions occur, and consequently yield higher recognition accuracy. This approach is similar to the one used in SPLICE [3].

In Figure 3 we present the same situation as in Figure 2, but now with the proposed approach. If we double the number of VQ clusters in the noisy speech feature domain, we can assign one VQ cluster and the corresponding Gaussian density to each combination of speech and noise.

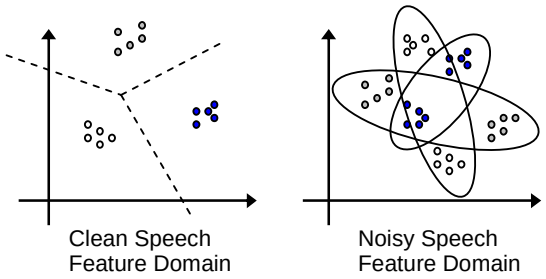

Figure 2: VQ Clustering in Clean Speech Feature Domain in Nonstationary Noise

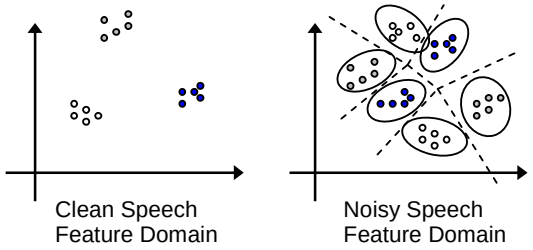

Figure 3: Proposed Approach: VQ Clustering in Noisy Speech Feature Domain in Nonstationary Noise

One important aspect of the proposed approach is that the number of clusters must be increased in order to capture all the possible combinations of speech and noise. Additionally, the stereo database for POF training must be increased significantly to enable a good coverage of the possible combinations of speech and noise to be observed.

Regarding the mathematical formulation of the clean feature POF estimate, the only difference between the old and new approaches is in the computation of the prior probabilities of the VQ regions. In the old approach it was the prior probability of the clean speech feature VQ clusters. In the new approach it is the prior probability of the noisy speech feature VQ clusters. However, the VQ clusters, now defined on the noisy speech space are substantially different from those obtained in the previous approach.

4. Comparing New and Old POF in Stationary and Nonstationary Noises

We compared the new and old POF algorithms in a command and control task [1]. Noises from a HMMWV vehicle (stationary noise), an M113 tank (stationary noise), and gunfire (nonstationary noise) were added at different levels of SNR, from 0 to 20 dB, to a clean speech database of 1580 utterances produced by five speakers. SNR-dependent POF mappings were trained. For POF training a separate database of 5000 phonetically varied sentences was used from the Wall Street Journal database. The size of the POF mappings was five feature vectors in the tapped-delay line and 300 Gaussians. The same POF size was used with the new and old POF algorithms. Recognition experiments were conducted on the noisy data at the different SNRs using three recognition systems: (1) a baseline system using the DynaSpeak recognizer [6] with 16 kHz bandwidth standard acoustic models, and no noise-robust processing, called the no compensation system, (2) the DynaSpeak system using the old POF algorithm trained using the stereo database and the specific noise data, and (3) same as (2) but with the new version of the POF algorithm.

Table 1: WER Results for Old POF and New POF Algorithms in Noisy Speech Corrupted with **M113 Noise**

SNR dB	No Comp	Old POF M113	New POF M113
20	1.04	0.91	0.91
15	2.26	1.22	1.24
10	10.93	2.86	2.59
5	43.57	7.63	6.12
0	83.76	24.71	19.15

Table 2: WER Results for Old POF and New POF Algorithms in Noisy Speech Corrupted with **HMMWV Noise**

SNR dB	No Comp	Old POF HMMWV	New POF HMMWV
20	1.93	1.06	1.18
15	5.33	1.93	1.93
10	21.64	4.11	3.57
5	60.39	10.87	9.73
0	89.83	30.98	25.95

The results in Tables 1 and 2 shows that in the no compensation system there is an important word error rate (WER) degradation as the SNR is reduced for both noises. By using the old POF algorithm a significant WER reduction is achieved compared to the no compensation case. The old POF algorithm performs on average 57.1% relative better in the M113 noise and on average 67.5% relative better in the HMMWV noise than the no compensation case. The new POF algorithm improves over the old POF algorithm mostly in the low SNR portion. The new POF algorithm performs on average 10% relative better in the M113 noise and 5.7% relative better in HMMWV noise than the original POF formulation. Also, the new POF algorithm performs on average 59.4% relative better in the M113 noise and on average 68.2% relative better in the HMMWV noise than the no compensation case.

Table 3: WER Results for Old POF and New POF Algorithms in Noisy Speech Corrupted with **Gunfire Noise**

SNR dB	No Comp	Old POF Gunfire	New POF Gunfire
20	1.70	1.06	1.06
15	4.21	2.24	2.05
10	11.58	6.16	5.21
5	34.81	19.11	13.80
0	61.04	46.00	32.05

The results in Table 3 show that in the no compensation system there is an important WER degradation as the SNR is reduced for the gunfire noise. By using the old POF algorithm a significant WER reduction is achieved compared to the no compensation case. The old POF algorithm performs on average 40.2% relative better in the gunfire noise than the no compensation case. The new POF results further reduce the WER compared to the old POF case. The new POF algorithm benefited from better clusters in the noisy feature domain and better transformations. The new POF algorithm performs on average 16% relative better than original POF formulation in the gunfire noise. Also, the new POF algorithm performs on average 50.3% relative better in the gunfire noise than the no compensation case.

Comparing the results of Tables 3 and Tables 1 and 2, the gain achieved by the new POF algorithm over the old POF algorithm is higher in the nonstationary noise than in the stationary noise.

5. Multinoise POF Feature Compensation

We used in this experiment a POF compensation system trained with multiple noises, both stationary and nonstationary, with the POF compensation trained in the matched testset noise only. The idea is to see how this compensation system generalizes to unknown stationary and nonstationary noises. The noises used in POF training were vehicle noises (HMMWV, Stryker), tank noises (e.g. M113, M109), gunfire noises, factory noise, and babble noise. In total there were ten different noises for POF training. Some of the noises were extracted from the NOISEX-92 noise database [7].

In Tables 4, 5, and 6 we compare several compensation systems in noisy speech with M113 vehicle noise, HMMWV vehicle noise, and gunfire noise. The noise was added at different SNR levels to the clean waveforms. In all cases the SNR was assumed known. In the first column the no compensation results are presented. The second column corresponds to the POF compensation trained with noisy speech using only the testing noise. The third column corresponds to the POF compensation trained with noisy speech with multiple noises, including the testing noise, resulting in a total of ten noises. Finally, the fourth column corresponds to the POF compensation trained with noisy speech with multiple noises, excluding the testing noise, resulting in a total of nine noises. The idea of this last compensation system is to see how the multinoise POF mapping generalizes to unknown stationary and nonstationary noises.

Table 4: WER Results for New POF Compensation, Multinoise With and Without Testing Noise in Noisy Speech corrupted by **M113 Noise**

SNR dB	No Comp	New POF M113 Noise	New POF Multinoise with M113	New POF Multinoise no M113
20	1.04	0.91	1.08	1.06
15	2.26	1.24	1.43	1.39
10	10.93	2.59	4.02	4.17
5	43.57	6.12	11.04	11.62
0	83.76	19.15	35.66	37.08

Table 5: WER Results for New POF Compensation, Multinoise With and Without Testing Noise in Noisy Speech corrupted by **HMMWV Noise**

SNR dB	No Comp	New POF HMMWV Noise	New POF Multinoise with HMMWV	New POF Multinoise no HMMWV
20	1.93	1.18	1.47	1.58
15	5.33	1.93	2.24	2.53
10	21.64	3.57	5.31	5.48
5	60.39	9.73	13.76	14.85
0	89.83	25.95	36.93	38.71

Table 6: WER Results for New POF Compensation, Multinoise With and Without Testing Noise in Noisy Speech corrupted by **Gunfire Noise**

SNR dB	No Comp	New POF Gunfire Noise	New POF Multinoise with Gunfire	New POF Multinoise no Gunfire
20	1.70	1.06	1.24	1.41
15	4.21	2.05	2.55	3.01
10	11.58	5.21	6.08	7.07
5	34.81	13.80	16.02	18.63
0	61.04	32.05	35.79	39.71

The results in Tables 4, 5, and 6 show that in the no compensation system there is an important WER degradation as the SNR is reduced for all the noises. The new POF algorithm obtains a significant WER reduction over the no compensation case since it is trained with the matched noise only. The multinoise POF with the testing noise still produces an important WER reduction over the no compensation case, but this reduction is not as significant as using a POF compensation system trained with using only the matched noise. The average WER reduction for the multinoise POF including the testing noise over the no compensation case is 45.6% for the M113 noise, 58.6% for the HMMWV noise, and 41.8% for the gunfire noise. The multinoise POF without the testing noise has a similar performance as the multinoise POF with the testing noise, which shows that it achieves a good level of generalization to the unknown noises. The average WER reduction for the multinoise POF with no testing noise over the no compensation case is 45.5% for the M113 noise, 55.5% for the HMMWV noise, and 33.1% for the gunfire noise.

6. Conclusions

We have proposed an extension of the POF algorithm that allows different combinations of speech and noise to have speech and noise combination-specific mappings. We showed that the proposed approach improves over the previous method, and that the gain is higher for nonstationary noises. The proposed approach also enabled us to train POF maps with multiple noises. We evaluated POF maps where the testing noise was included as one of the multiple noises in training and where the testing noise was not included as one of the multiple noises in training. Both maps still achieved a substantial WER reduction over the no compensation case.

7. Acknowledgments

This research was in part funded by the U.S. Army in the "Collaborative Alliance for Robotics Technology" project under prime contract DAAD19-01-2-0012, subcontract 9613P with General Dynamics Robotic Systems. The views herein are those of the authors and should not be interpreted as representing the policies of the funding agency or prime contractor.

8. References

[1] G. Myers, H. Franco, M. Graciarena, C. Cowan, F. Cesari, and V. Abrash, "*Noise-Robust Spoken Language Interface*", in CTA 2003 Symposium on Robotics, Adelphi, MD, pp. 135-139, April 2003.

[2] L. Neumeyer and M. Weintraub, "*Probabilistic Optimum Filtering for Robust Speech Recognition*", in Proc. Intl. Conf. on Acoustics, Speech and Signal Processing'94, Adelaide, Australia, 1994, pp. I417-I420.

[3] J. Droppo, L. Deng and A. Acero, "*Evaluation of the SPLICE Algorithm on the Aurora2 Database*", pp. 217-220, Proc. Eurospeech, Aalborg, Denmark 2001.

[4] L. Buera Rodríguez, E. Lleida Solano, A. Miguel Artiaga, and A. Ortega Gimenez. "*Multi-Environments Model Based Linear Normalization for Speech Recognition in Car Conditions*", International Conference on Audio, Speech and Signal Processing, ICASSP-2004, Montreal, Canada, May 2004.

[5] Y. Linde, A. Buzo, and R.M. Gray, "*An Algorithm for Vector Quantizer Design*," IEEE Transactions in Communications, vol. 28, pp. 84-95, January 1980.

[6] H. Franco, J. Zheng, J. Butzberger, F. Cesari, M. Frandsen, J. Arnold, R. Rao, A. Stolcke, and V. Abrash, "*DynaSpeak: SRI's Scalable Speech Recognizer for Embedded and Mobile Systems*", Proc. Human Language Technology Conference (HLT-2002), San Diego, CA, 2002.

[7] NOISEX database samples available at http://spib.rice.edu/spib/select_noise.html